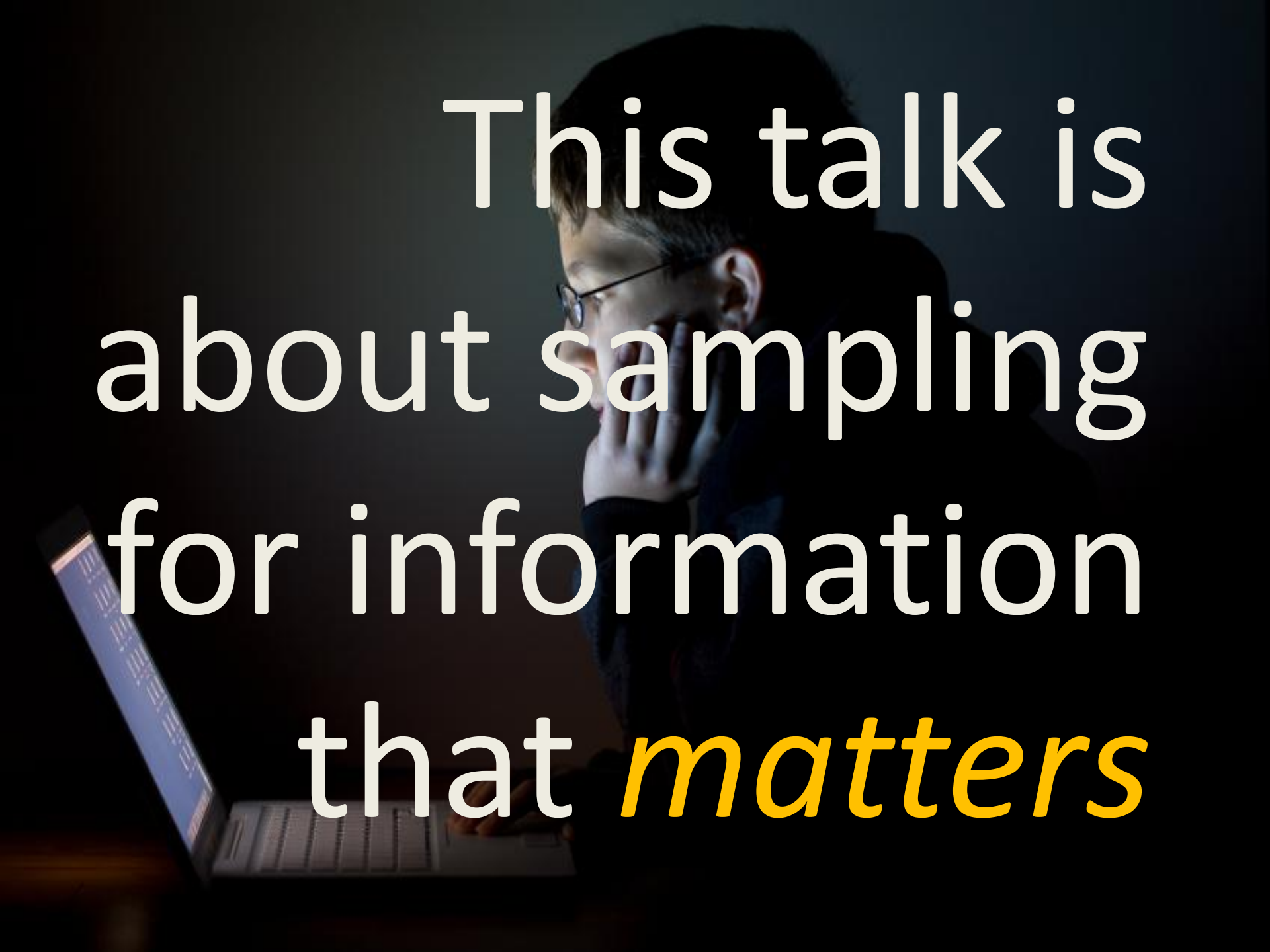
Information That Matters: Investigating Relevance in Social Communication Media

Munmun De Choudhury
PhD Candidate, Computer Science
Arizona State University
http://www.public.asu.edu/~mdechoud/


This talk is
about sampling
for information
that *matters*

Modern Social Interactional Modes

140 characters can cause revolutions

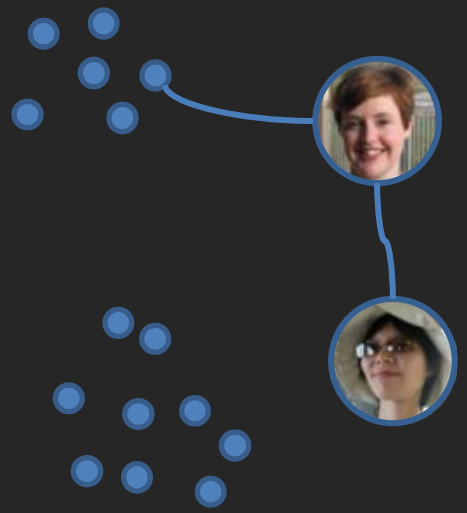

PINPOINT NEWS TRACKER
EARTHQUAKE
4.0 MAGNITUDE
Microsoft Virtual Earth
© 2007 Microsoft Corporation
© Harris Corp, Earthstar Geographics
La Jolla
PACIFIC OCEAN
Pacific Beach
Ocean Beach


But the social web is changing at a fast rate
HEALTH MATTERS

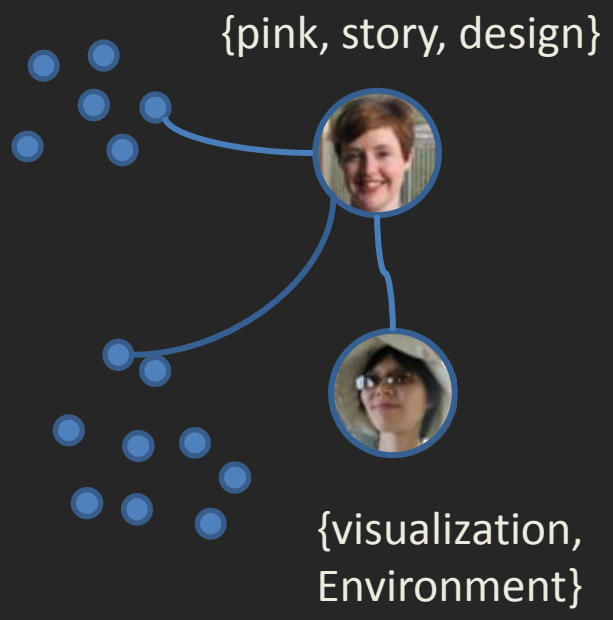
And
what exactly
is changing?

New people appear

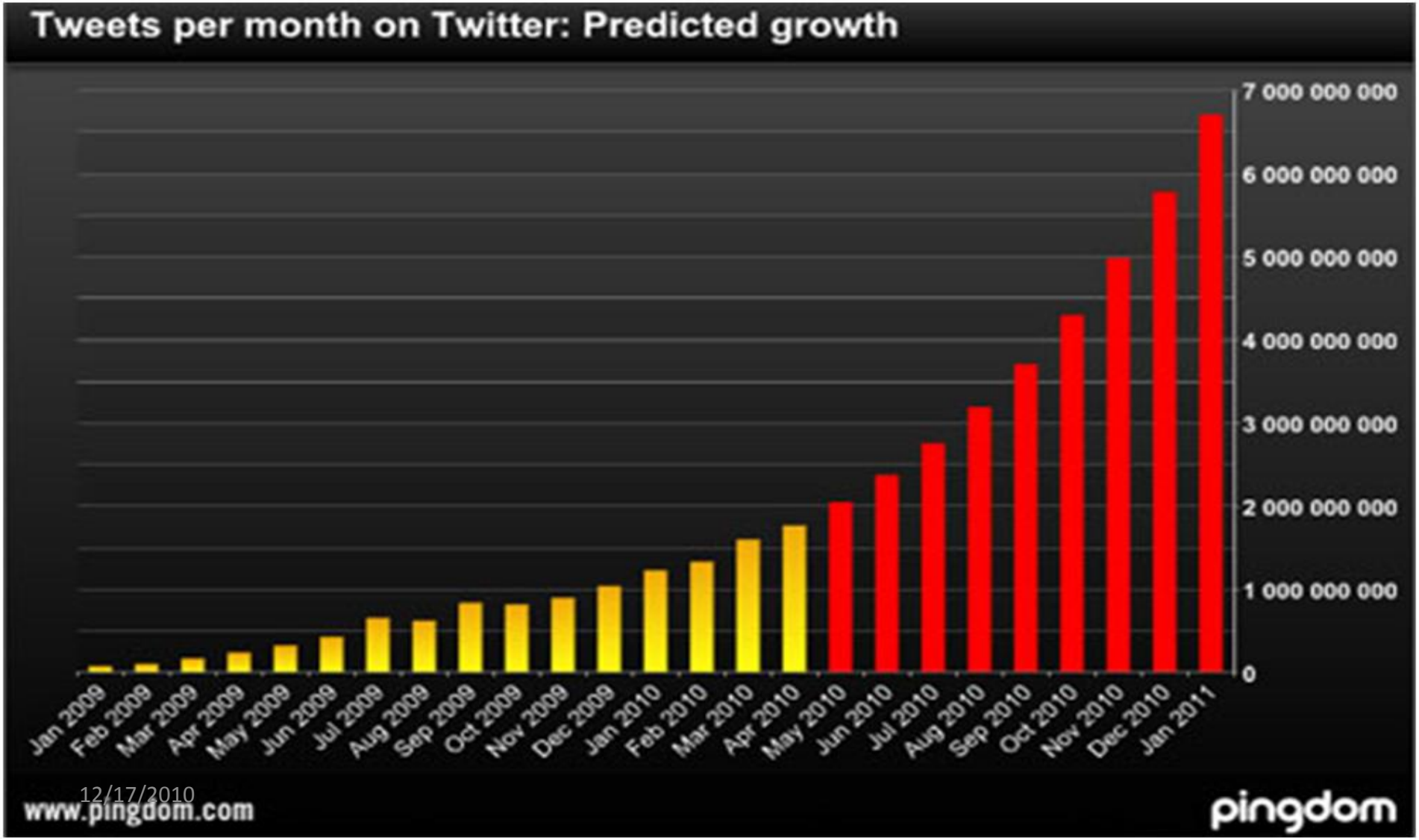
New ties are formed



New interactional data appears too!



By April 2010, http://www.twitter.com/ was receiving over 600 million search queries per day (Huffington Post).



Facebook hits milestone: Half a billion users

Social networking site marks occasion with new Facebook Stories app, interview with Diane Sawyer

By Sharon Gaudin

July 21, 2010 02:29 PM ET

Comments (8) Recommended (16) Share

We are attracted to social media, in part due to large scale datasets


Is there
something
more
fundamental
happening here
than just scale?

Two simple questions

How do we infer meaningful human networks?

How do we identify valuable social media content?

Question I

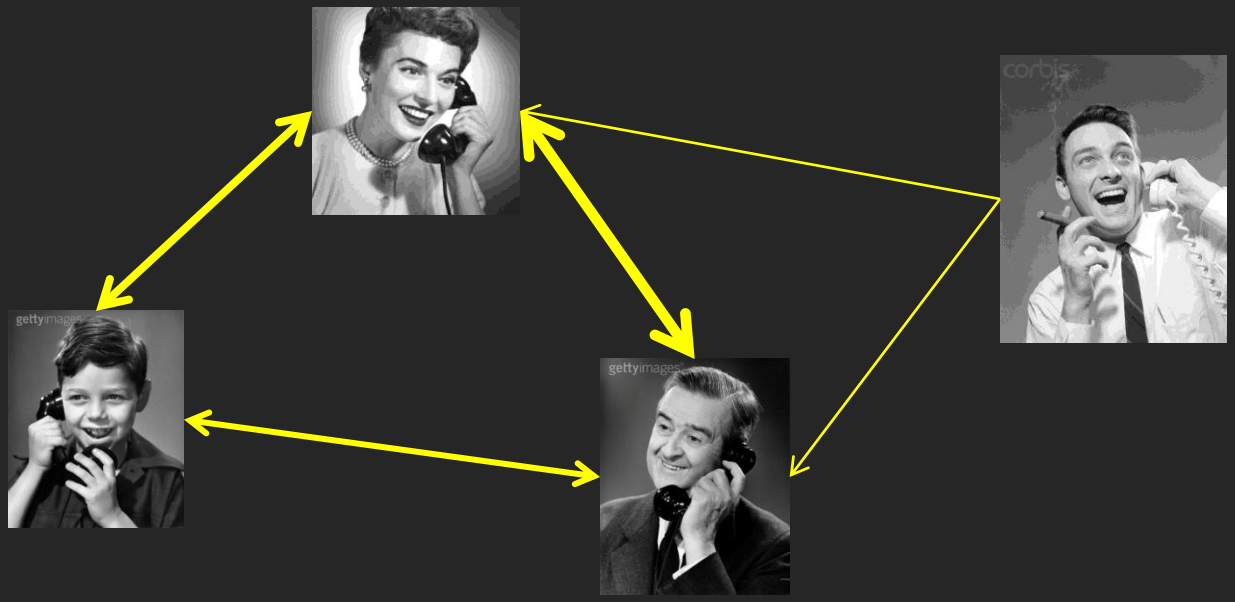
- With Winter Mason, Jake Hofman and Duncan Watts, during internship at Yahoo! Research, summer 2009
  - Published at Intl. Conference on World Wide Web (WWW) 2010

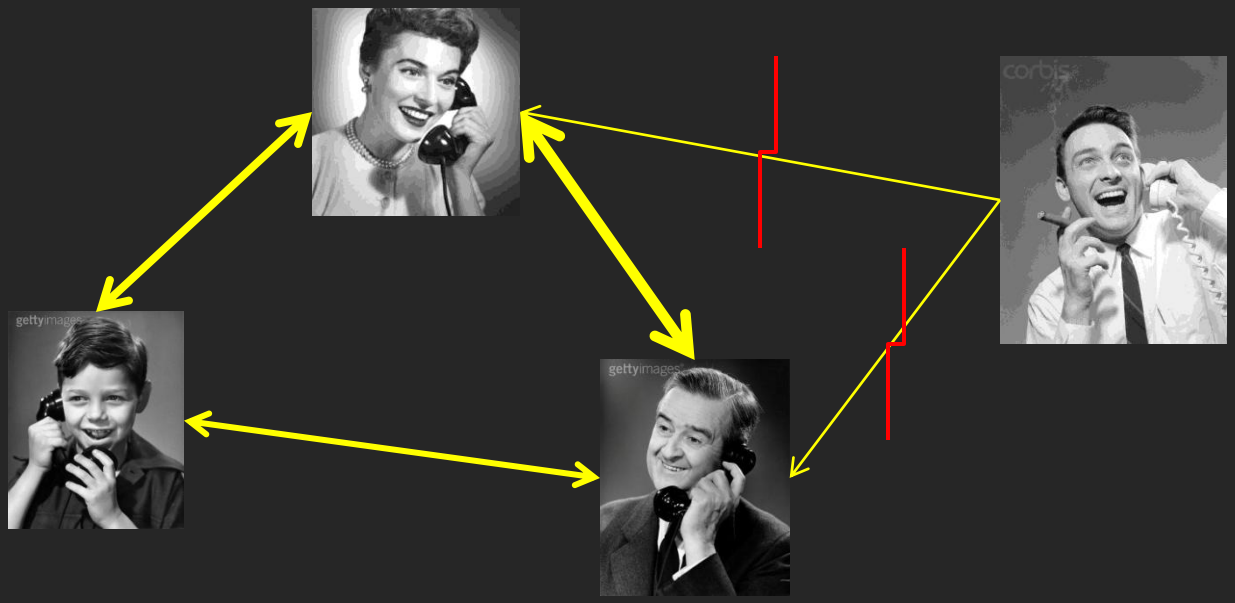
How to choose a relevant tie?

How frequently do you talk to your “best friend”?

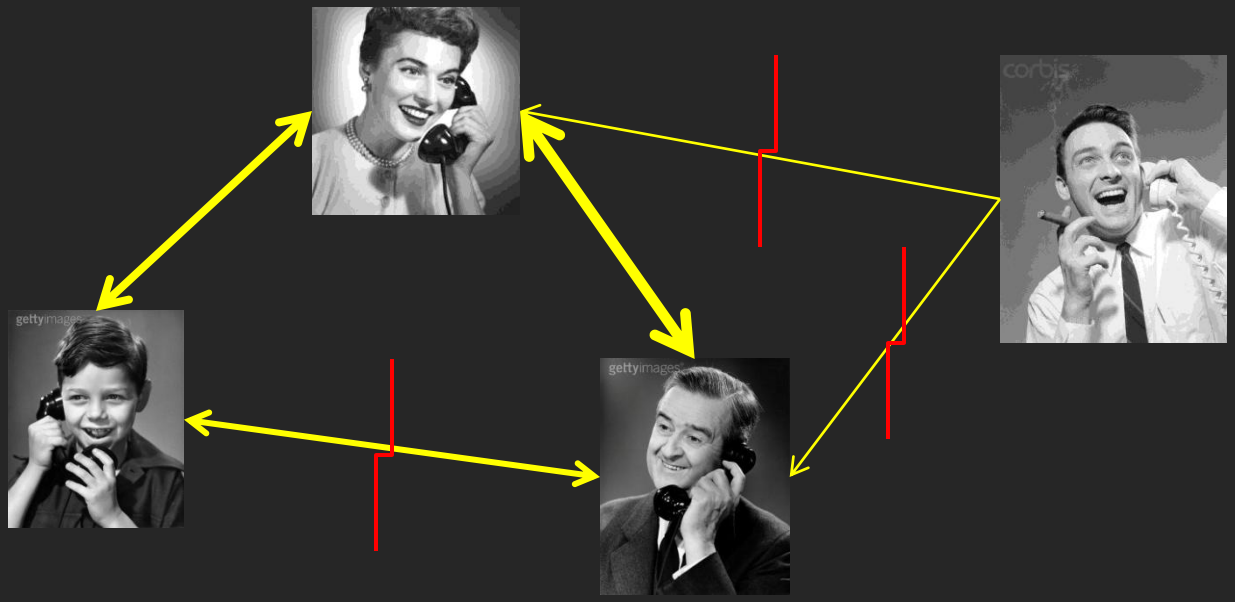
Social ties from communication data

- Reasonable definitions:
  - At least one communication in past year
  - Average of one communication every week
  - One reciprocated communication in past month
- What is the research question?
  - Search on network
  - Information diffusion
  - Uncovering hidden node properties
- *One method: define a minimum threshold*

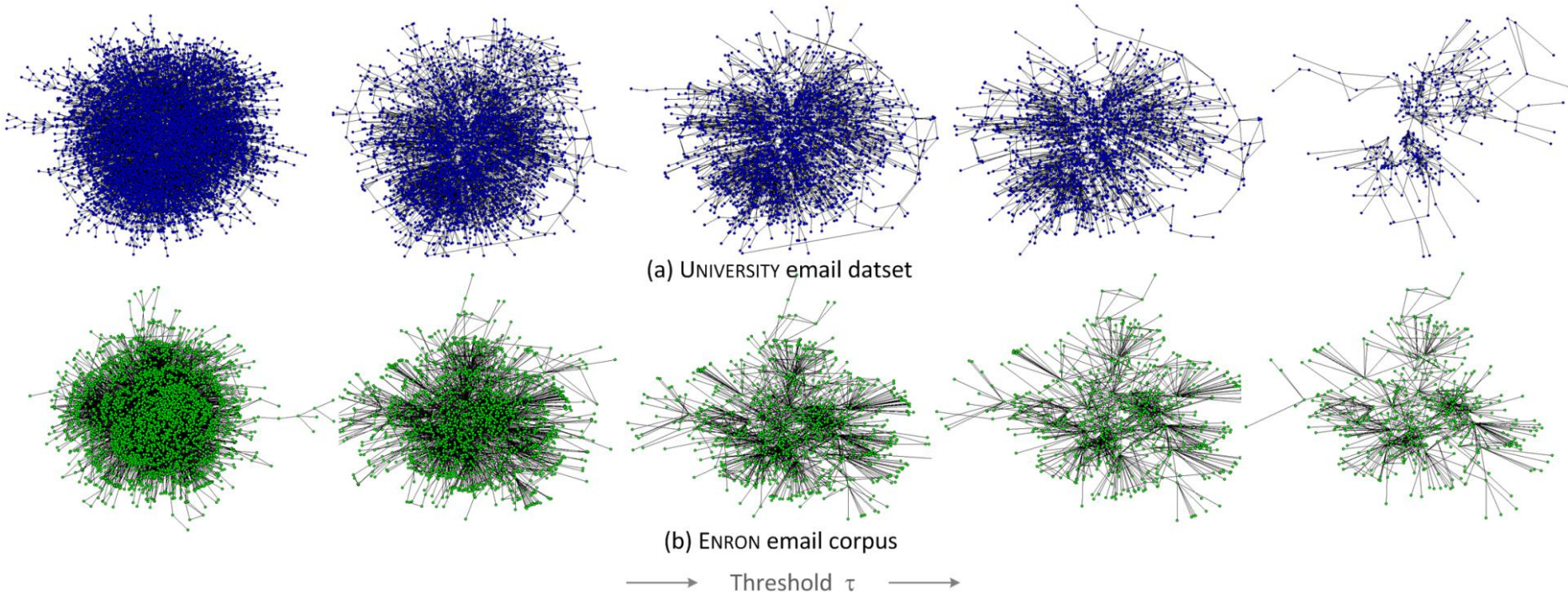

gettyimages
gettyimages
corbis



Threshold =



Threshold =



Why inferring the relevant tie matters

Our Contributions

- **Goal:**
  - Infer networks for various definitions of “threshold” over a tie
  - Study the impact of different thresholded networks on:
    - descriptive statistics and
    - ability of the network in predicting node characteristics
- **Outcomes:**
  - There exists a non-trivial threshold on edge weights over which our set of chosen prediction tasks seem to yield maximum accuracy
  - The optimal range of threshold values appears to be relatively consistent across datasets and prediction tasks

University Email

1. a complied registry of all email (incoming and outgoing, as recorded in server logs) associated with individuals at a large university in the US, comprising undergraduate and graduate students, faculty, and staff
2. Focus on a consistent user set across all semesters - 19,817 individuals
3. 1.09M emails; disregard emails involving non-university domain
4. PS: content of emails not available

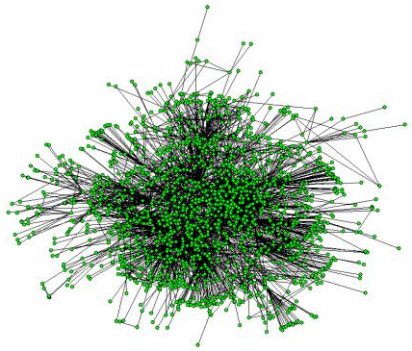
Enron Corporate Email

1. a repository of the emails exchanged internally among the employees at the Enron Corporation, obtained through a subpoena as part of an investigation by the Federal Energy Regulatory Commission (FERC) and then made public
2. 4,736 individuals
3. 1.06M emails
4. 4 years (1998-2002)

"Thresholded" Networks

- **Edge definition:**
  - Symmetric edge based on the frequency of email communication
  - Geometric mean of the annualized rate of messages exchanged over the span of two and four years respectively. For users u_i and u_j:

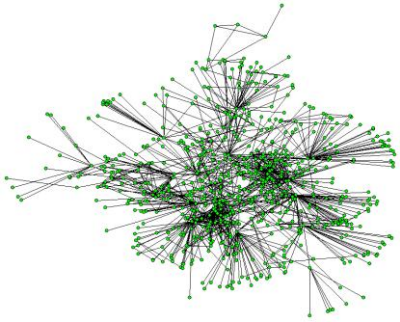
$$e_{ij} = \sqrt{w_{ij} w_{ji}}$$



τ=5 emails per year

- **Edge threshold:**
  - Minimum of τ emails between each pair of individuals, over a period of time T. Hence we construct the social graph $G(V,E;\tau)$ such that,

$$e_{ij} \in E \text{ if and only if, } e_{ij} \geq \frac{\tau}{T}.$$



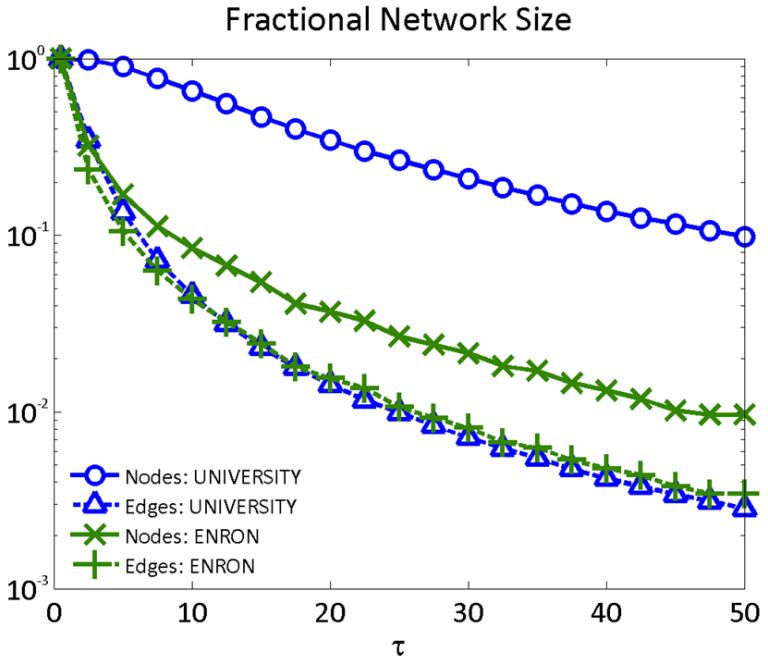
τ=15 emails per year

- Family of networks: $\{G(\tau_1), G(\tau_2), \ldots, G(\tau_K)\}$

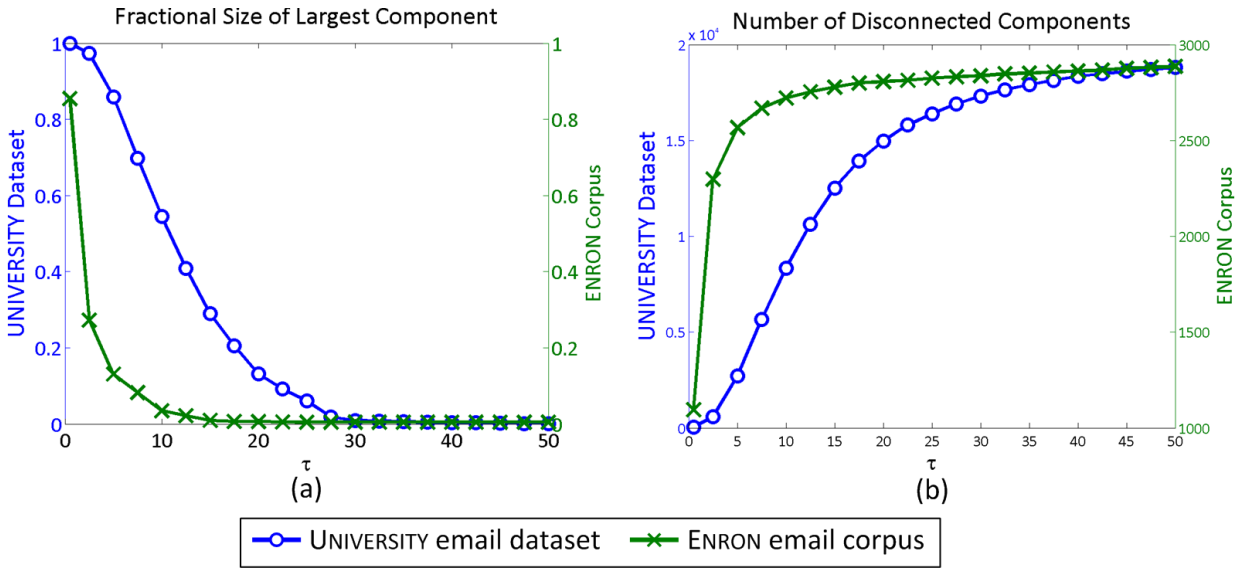
Network Descriptive Statistics

Network-level Features

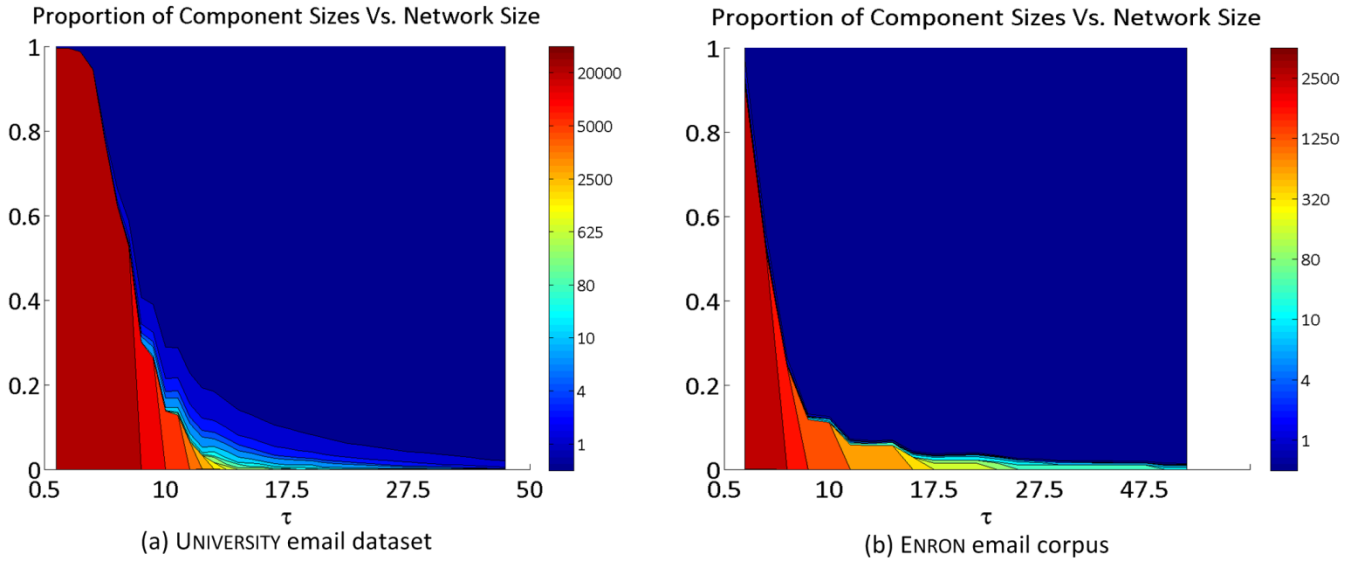
- Vary threshold τbetween 0.5 and 50 emails per year.
  - The natural starting point for τ is the lowest value for which both networks are defined. It is $\tau = 0.5$, or one email over two years for the University dataset
- Network density:
  - Number of edges in both datasets decrease rapidly as threshold increases
  - Number of nodes decreases for both, though more rapidly for Enron dataset



Number of Components

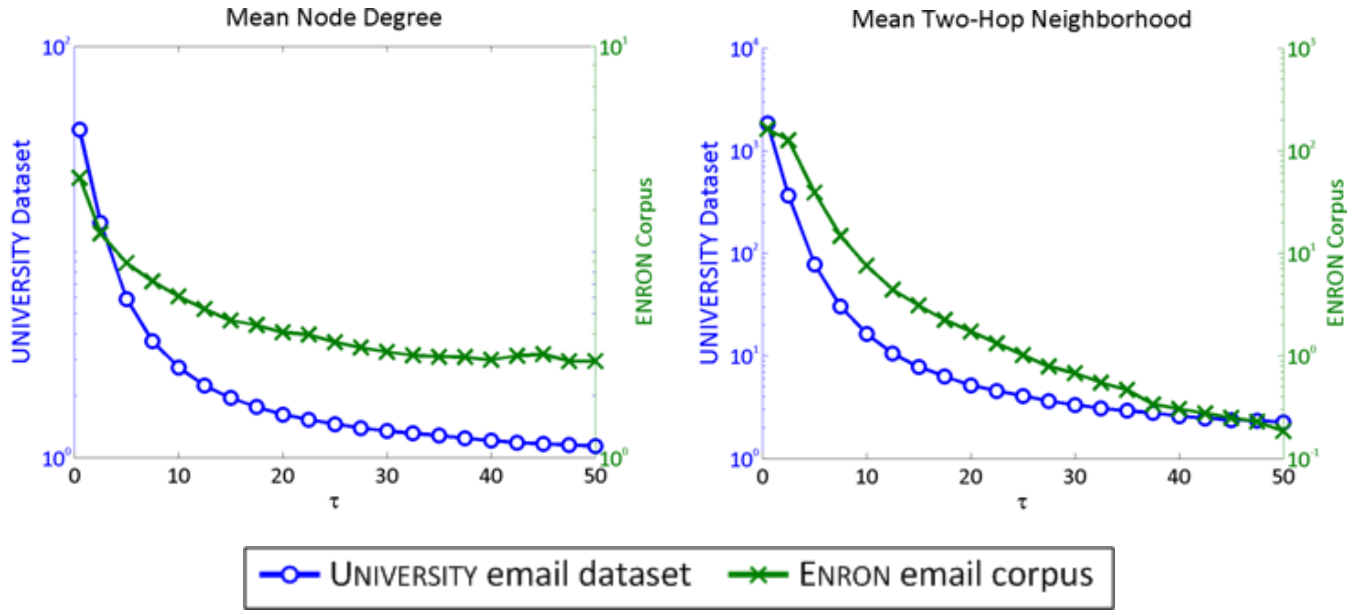


Relative Sizes of Components



Node-level Features

- *Reach* of a node:
  - Node degree
  - Average neighbor degree
  - Size of two-hop neighborhood
    - the count of all of the node's neighbors plus all of the node's neighbor's neighbors



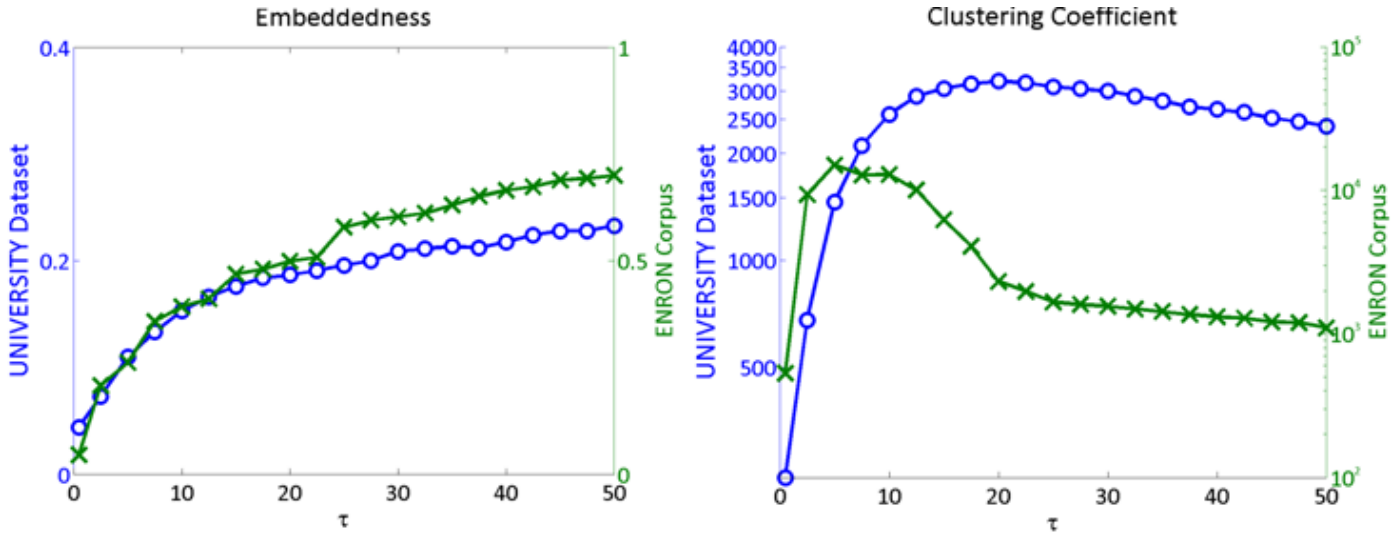
Node-level Features (Contd.)

- *Closure* of the ego-network:
  - Embeddedness

$$\Phi_i = \frac{1}{k_i} \sum_{u_j \in \Gamma_i} \frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|}, \text{ where } \Gamma_i = \{u_j : e_{ij} \in E\} \text{ and } k_i = |\Gamma_i|.$$

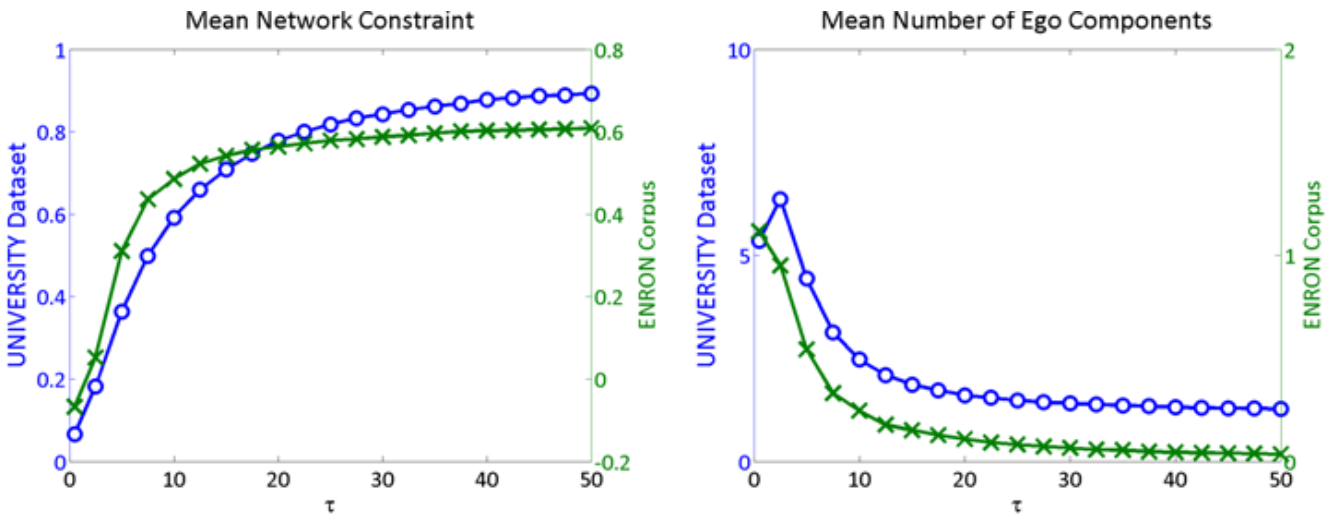
  - Normalized clustering coefficient

$$C_i = \frac{c_i}{k_i(N-1)}, \text{ where the clustering coefficient, } c_i = \frac{2|e_{jm}|}{k_i(k_i-1)}.$$



Node-level Features (Contd.)

- To what extent does a node *bridge* communities:
  - Network constraint
  - Number of ego components
    - count of the number of connected components that remain when the focal node and its incident edges are removed



How to choose the *right* threshold?

Premise of Prediction

Define an edge according to the research problem of interest…

- Making predictions – University dataset
  - Node Status (“faculty”, “student”, etc.)
  - Gender
  - Future communication activity i.e. the number of emails sent by a user at a future time slice
  - Community detection (school assignment)
- Making predictions – Enron dataset
  - Node Status (“Director”, “Manager”, etc.)
  - Future communication activity

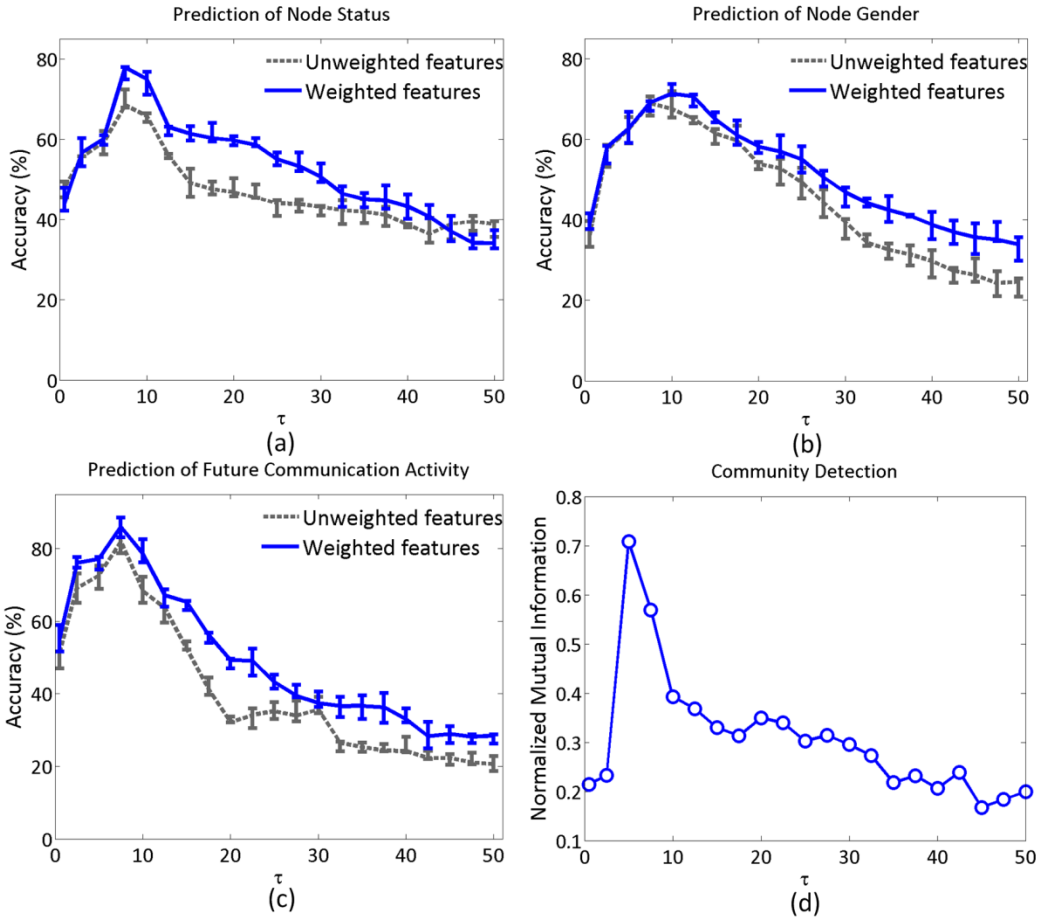
Prediction Task: Node Status/Gender

Prediction Task: Future Communication

Prediction Task: Community Detection

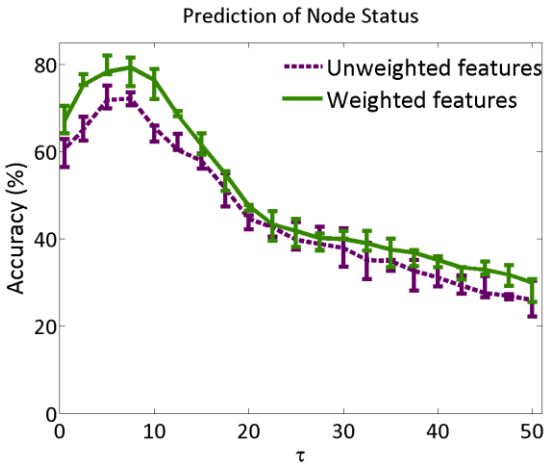
Experimental Results (University Email)

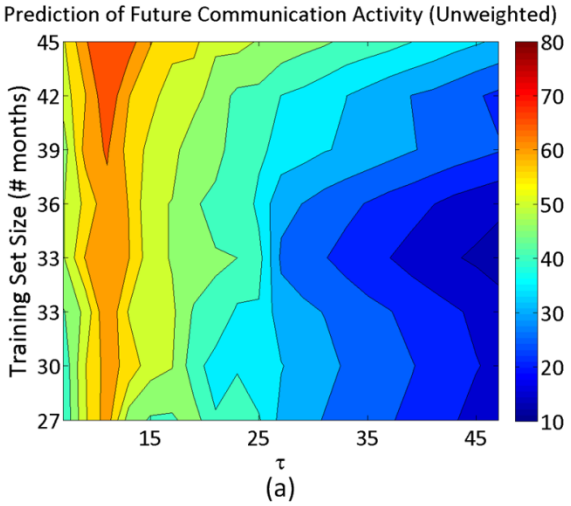
- Peak accuracy in different prediction tasks occurs at a non-trivial τ.
- There is ~30% boost in accuracy over unthresholded network.

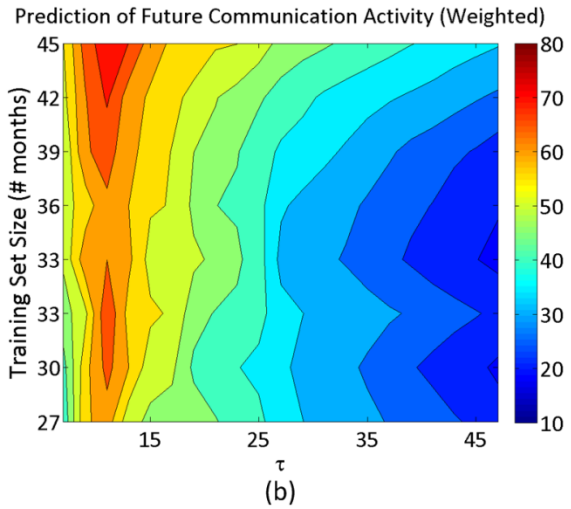


Experimental Results (Enron Email)

- Best accuracy occurs at τ=7.5 for the two prediction tasks
- Accuracy increases from ~60% to ~70% from unthresholded graph to optimal τ for unweighted features, and ~65% to ~80% for weighted features







Observations

- Finding Optimal Threshold
  - Accuracy maximized at non-obvious point
  - Increase in accuracy from unthresholded graph as much as **~30%**
  - Increase in accuracy exists even including information about weights at edges; therefore **deleting edges removes noise** (increasing signal)

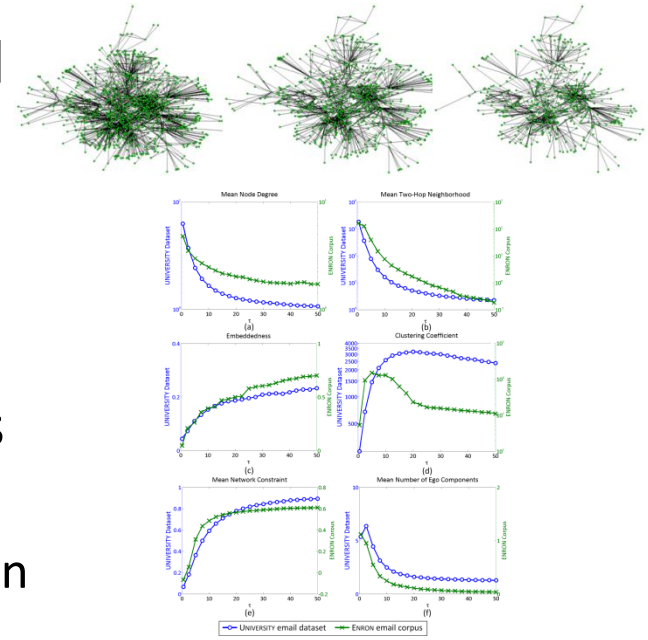
- Optimal threshold at consistent value
  - For different prediction tasks
  - For different data sets

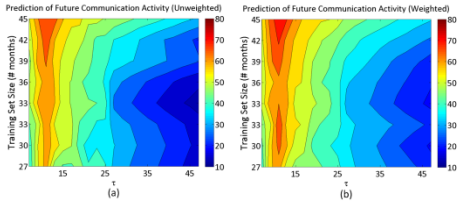
Discussion

- Initial assumptions made on social graph construction:
  - Elimination of out-of-network nodes, focusing on a consistent user set over time
  - Geometric mean: alternative definitions of an edge?
  - Considered symmetric edges: communication is often asymmetrical
- Only tested with email datasets
- Type of prediction tasks constrained by available data
- Thresholds on edge weights are not the only way to define edges

Conclusions

- Network analysis of communication data takes as input some set of observations and infers from these data a set of relations to which social and psychological meaning is attached
  - Network inference procedure largely ad-hoc
- We have addressed a narrow version of this general problem:
  - how to determine an optimal threshold condition for edges so as to predict particular node attributes (e.g. gender, status) or behavior
  - The prediction accuracies peak in a non-obvious—yet relatively narrow, threshold range across both datasets





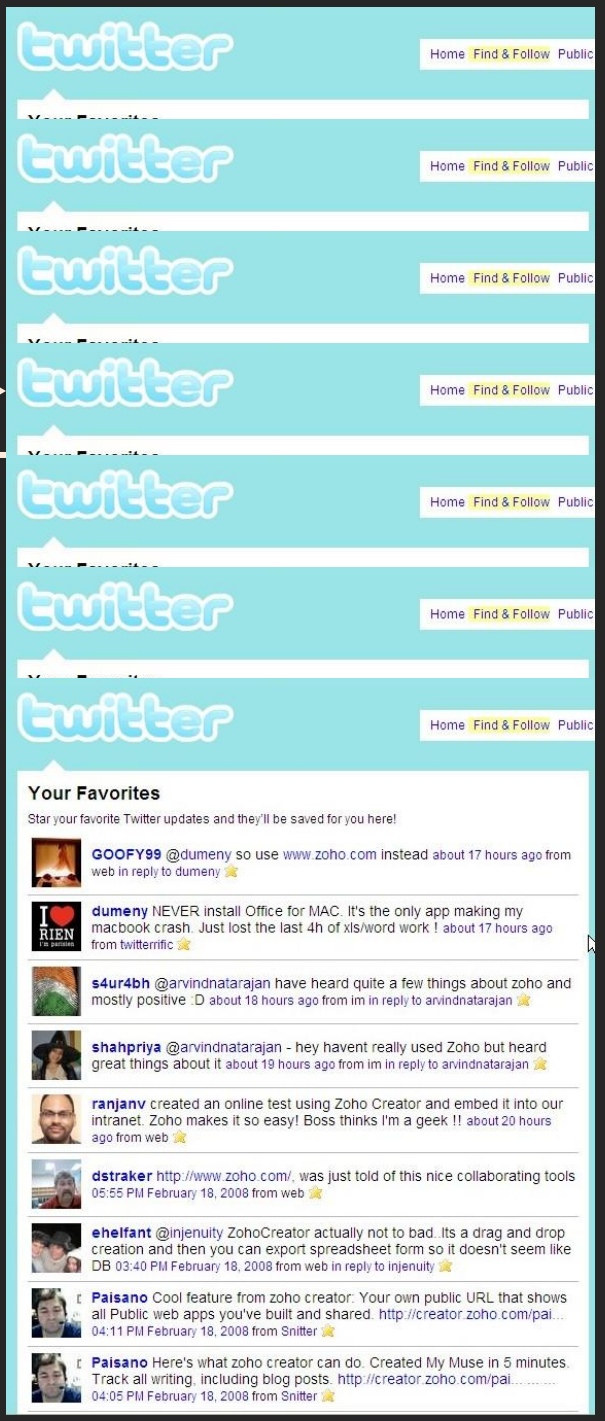
Incorporate model of tie relevance in prediction task?

Learn optimal threshold for known feature, test on unknown feature?

Question II

- With Scott Counts and Mary Czerwinski, during internship at Microsoft Research, summer 2010
  - Under review at a double-blind conference
  - Under review at Intl. Conference on World Wide Web (WWW) 2011
  - Imminent article to be submitted to the *Science* journal

Have you ever read your *entire* Twitter timeline carefully?

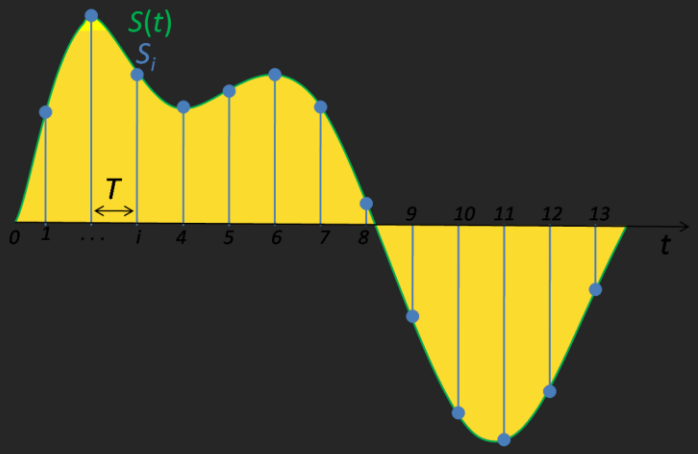


"Information overload" problem – *Get me the right content!*

How do we identify the most “relevant” or “”best” items on a topic, from millions and even billions of units of social media content?

Let’s contrast this with a familiar example

Discrete, regular and fixed sampling lattice



•Shannon-Nyquist sampling theorem: "If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $1/(2B)$ seconds apart."

Time to sample each pixel is constant

Note that the web activity has no notion of bandwidth!

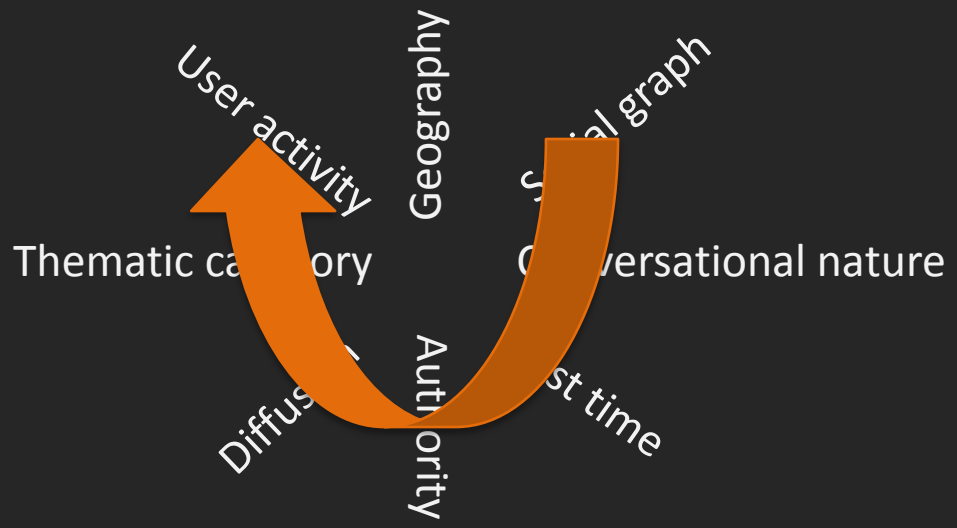
Interfaces / tools	#Responses
Twitter website	50
Twitter clients, such as Tweetdeck, Twitterific etc.	25
Search engines, such as Bing Social	19
Third party apps, such as Twitter plugin for Google	9

Unidimensionalinformation presentation; but social media information is diverse.

Social media sampling: consumption related to human cognition

Characteristics of social media – high dimensionality

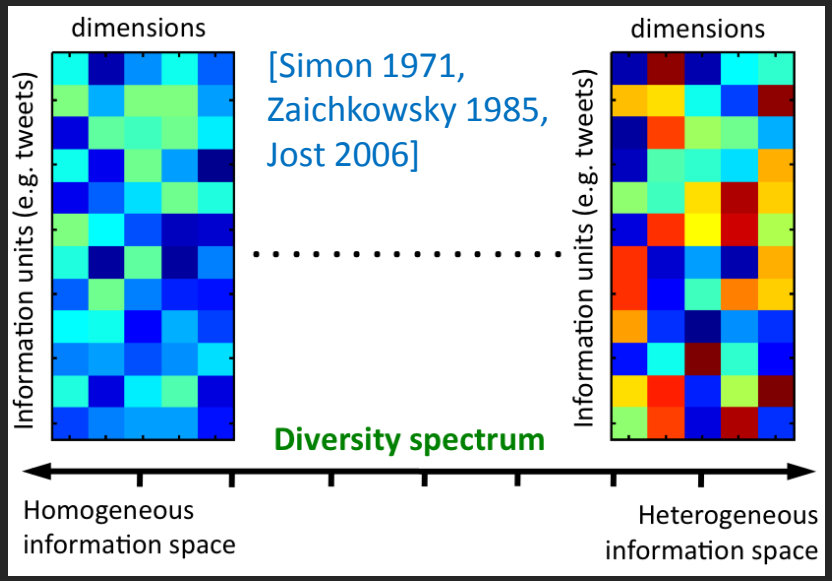
User cognition – mechanisms of human information processing



Information Diversity

Main Idea

Characteristics of social media – high dimensionality



Information Diversity

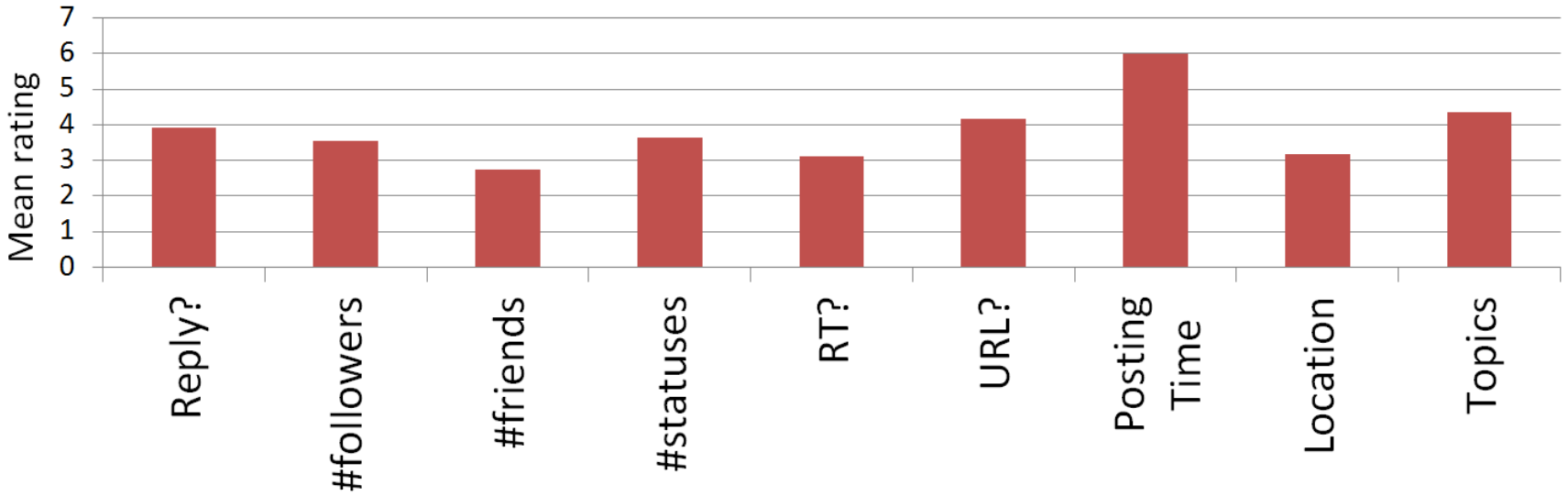
User cognition – mechanisms of human information processing

- Engagement
- Memory encoding
- Interestingness
- Informativeness

Main Idea

Dimensional Importance

- Survey based feedback on the importance of different dimensions – referred to as “concentration parameters”.
  - Participants (11 ‘active’ Twitter users) were requested to rate each of the tweet dimensions on a scale of 1 through 7, where 1 implied “not important at all”, and 7 meant “highly important”.
  - The survey also allowed them to identify other dimensions that they might think to be significant.



Social media sampling

- Our solution is motivated by the work in the signal processing literature on "compressive sensing" [Candes 2006]:
  - Social media content over time can be considered as signals that often bear the property of being highly "sparse" [Romberg 2008].
  - Compressive sensing can be used to exploit this notion of sparsity in social media content based signals to describe it (i.e. a tweet stream) as a linear combination of a very small number of basis components.
- Given $\Psi \in R^{N \times K}$, we are interested in the "underdetermined" case $M << N$, M is the number of basis functions whose coefficients can reconstruct Ψ.
  - Formally, our goal is to find $\Psi_S \in R^{M \times K}$, i.e. the general problem of reconstructing $\Psi \in R^{N \times K}$ from linear measurements Ψ_S about Ψ of the form: $\Psi_S = \Phi\Psi$, Φ is the transformation matrix.

Social media sampling (contd.)

- We utilize the popular wavelet transform, called "Haar wavelet" for reconstruction of Φ.
- Given $\Psi \in R^{N \times K}$, a tweet $t_i \in \Psi$ can be written as $t_i = \Lambda^* f$, Λ is the $N \times N$ matrix, Λ^* is the ortho-normal basis.
  - Hence $t_{iS} = \Phi f$ or $t_{iS} = \Phi' t_i$, where $\Phi' = \Phi \Lambda^*$.
  - We could recover f by finding among all coefficients consistent with the data Ψ, the decomposition with minimum L_1-norm:

Social media sampling (contd.)

- We perform iterative clustering for tweet sample generation – based on entropy distortion minimization technique.
  - The samples are constructed given a sampling ratio ρ and a diversity parameter value ω.
  - The (sub)-optimal sample to be constructed is represented as, $\Psi_S^*(\rho, \omega)$.
- Start with a random tweet as a sample seed.
- Iteratively keep on adding tweets from Ψ_S, say t_i, such that the distortion (in terms of L_1-norm) of entropy of the sample (say, $\Psi_S(i,\omega)$) on addition of the tweet t_i is least with respect to the specified diversity measure ω.

$$\arg \min_{t_i \in \Psi_S, t_i \notin \Psi_S(i-1,\omega)} \left\| H_O(\Psi_S(i,\omega)) - \omega \right\|_{L_1}, \text{ where}$$

$$H_O(\Psi_S(i,\omega)) = -\sum_{k=1}^{K} P(\vec{t}_{ik}).\log P(\vec{t}_{ik}) / H_{\max},\ t_i \in \Psi_S \text{ and } H_{\max} = \ln K.$$

Sample ordering

- We present a simple entropy distortion based ordering technique of the tweets in the (sub) optimal sample $\Psi_S^*(\rho, \omega)$.
  - Our central intuition is that the ordering should be based on how close a particular tweet $t_i \in \Psi_S^*(\rho, \omega)$ is, in terms of its different sampling dimensions K, with respect to the specified diversity parameter ω.
  - Hence compute the entropy distortion:

$$\min_{t_i} \left\| H_O(\Psi_S^*(i, \omega)) - \omega \right\|_{L_1}$$

  - The lower the distortion, the higher is the "rank" / "position" of the tweet in the ordered sample.

How does this method compare to state-of-the-art techniques?

Twitter, **full fire-hose**, June 2010, total 1.4 Billion tweets

- (data processing on Microsoft COSMOS using SCOPE and DryadLINQ)

Qualitative evaluation

@Paramedic_Fla	Some oil spill events from Monday, June 7, 2010 http://bit.ly/cRwfXn
@miamiauto	Some oil spill events from Monday, June 7, 2010: A summary of events on Monday, June 7, Day 48 of the Gulf of Mexi... http://bit.ly/9HNG9Z
@franklanguage	RT @DAYLEE F@CK that! Broken pipe is not NATURAL! RT @RayBeckermanFreedomWorks CEO, Calls Oil Spill Natural Disaster http://bit.ly/coUY4l
@Teasdallqrb	Public offers 'helpful' ideas on containing BP oil spill - NEWS.com.au

[Twitter search-alike] Most Recent tweets

@_paigenesss	RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a
@LeiaOfAlderaan	Citizen Speaks The Truth ON BP Gulf Oil Spill--the Govt, BP Are Doing Nothing, There Are No Leaders Here http://bit.ly/BP-Gulf-Oil-Spill
@Faustinagwlxo	WOOW! NO WAY! so brutal! http://ilil.me/h MTV Movie Summer Jam WWDC Oil Spill Xtina Another Cinderella Story
@minxdeluxe	RT @OliBarrett: Visualizing the BP Oil Spill http://www.ifitwasmyhome.com/

[Bing-alike] Most tweeted URL-containing tweets

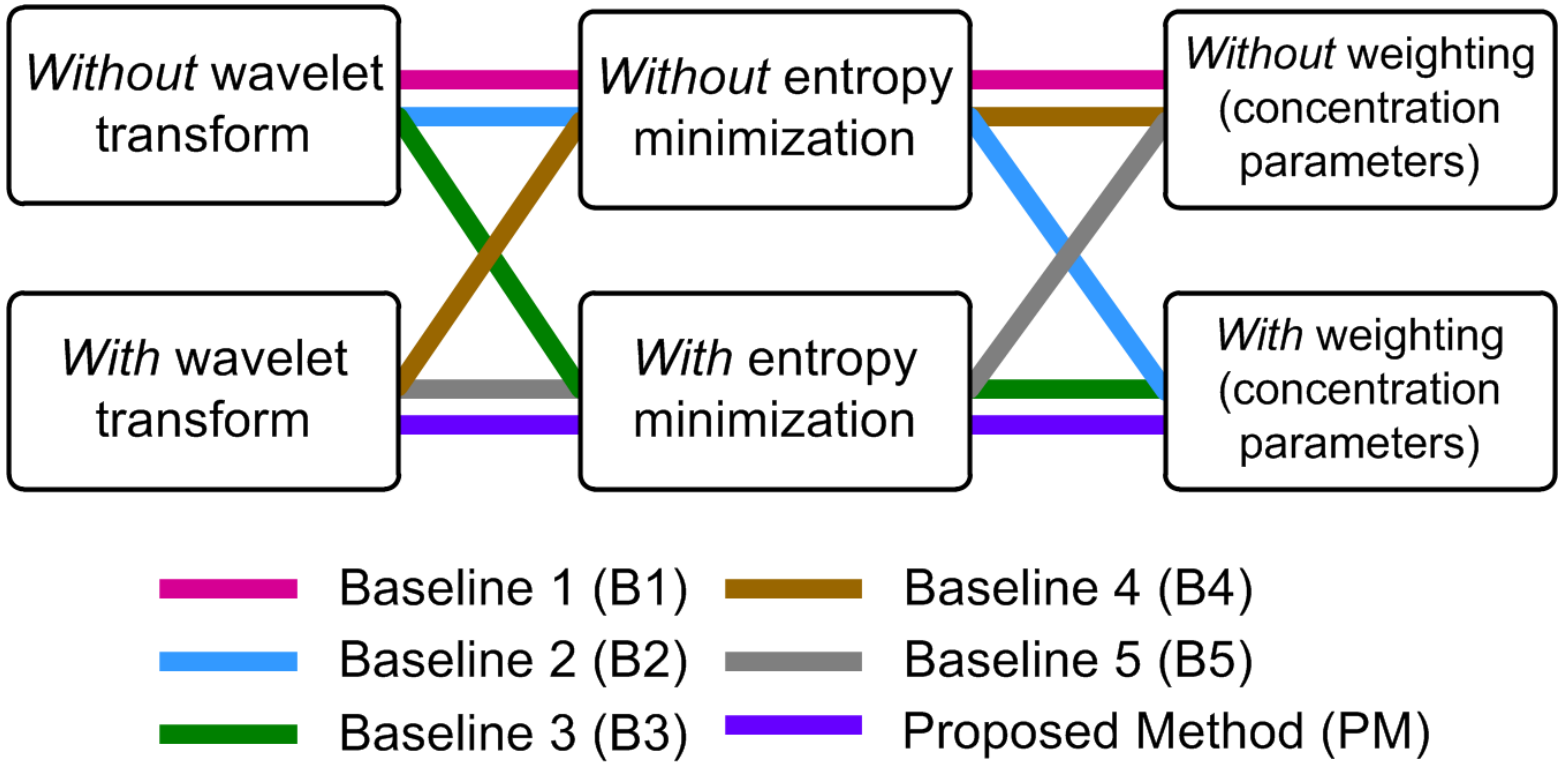
@JosephAGallant	Erin Brockovich to meet with fishermen who say oil spill dispersant used by BP made them sick. http://huff.to/aGVWIl #tcot #BP #oilspill
@dixie_patriot	Oil spill cap catching about 10,000 barrels a day|LONDON ? BP's oil spill cap, designed to stop a huge leak from .. http://oohja.com/xeWhD
@MoCuad	My heart breaks all over again, every time I'm reminded of the oil spill.
@NFGNL	Looking for Liability in BP's Gulf Oil Spill: White Collar Watch examines the potential criminal and civil liab.. http://nyti.ms/9lUMaT

Proposed Method (user-weighted; ω=0.1; ordered)

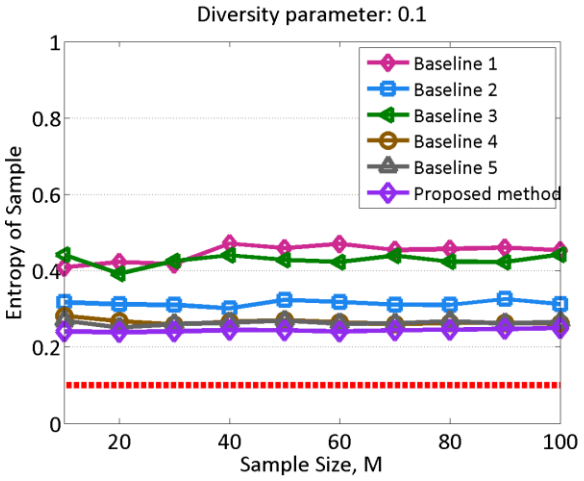
@jameelee	How You Can Volunteer to Clean Up the Gulf of Mexico Oil Spill http://ow.ly/1V3cu
@conchkid	Gulf;Oil Spill Many Federal Judges Have Links To Oil Industry http://bit.ly/9v45UT
@NewsOnGreen	BP Oil Spill: Containment Cap To Be Replaced Next Month http://dlvr.it/1WDZ8
@TrinitySaveNeo	Citizen Speaks The Truth ON BP Gulf Oil Spill--the Govt, BP Are Doing Nothing, There Are No Leaders Here http://bit.ly/BP-Gulf-Oil-Spill

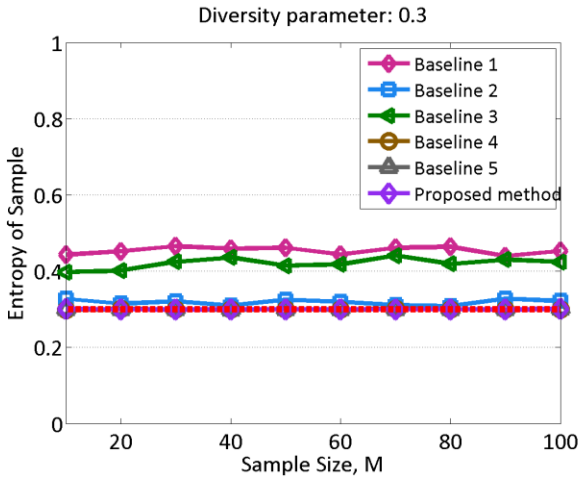
Proposed Method (user-weighted; ω=0.6; ordered)

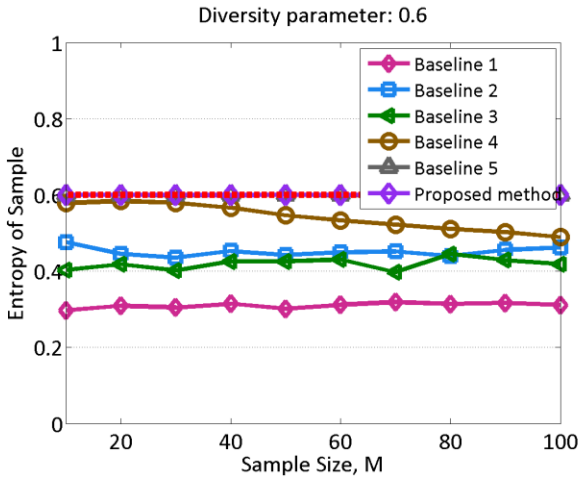
Quantitative evaluation framework



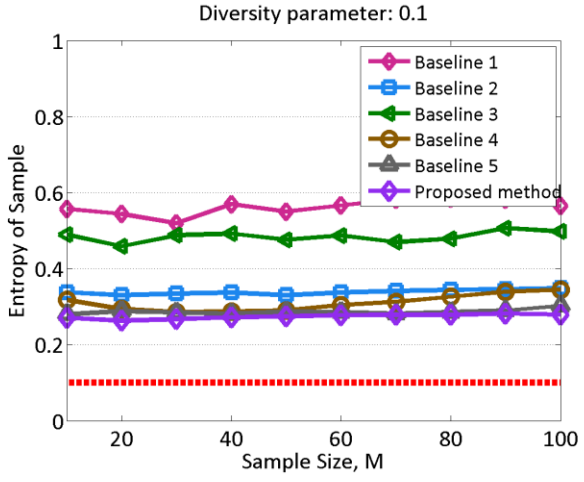
Quantitative evaluation

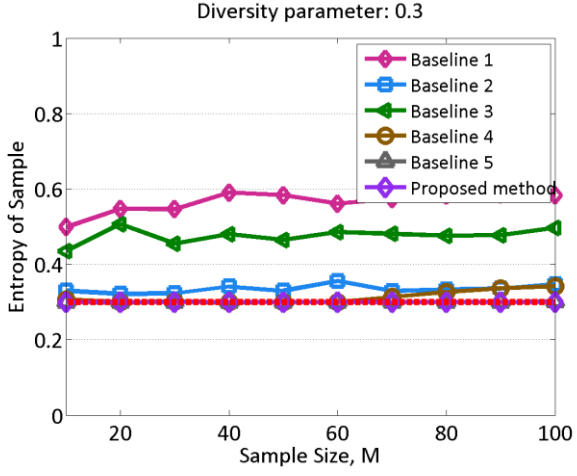


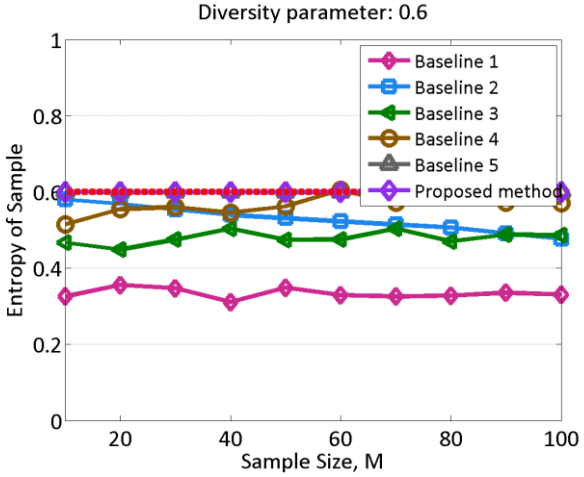




Oil Spill

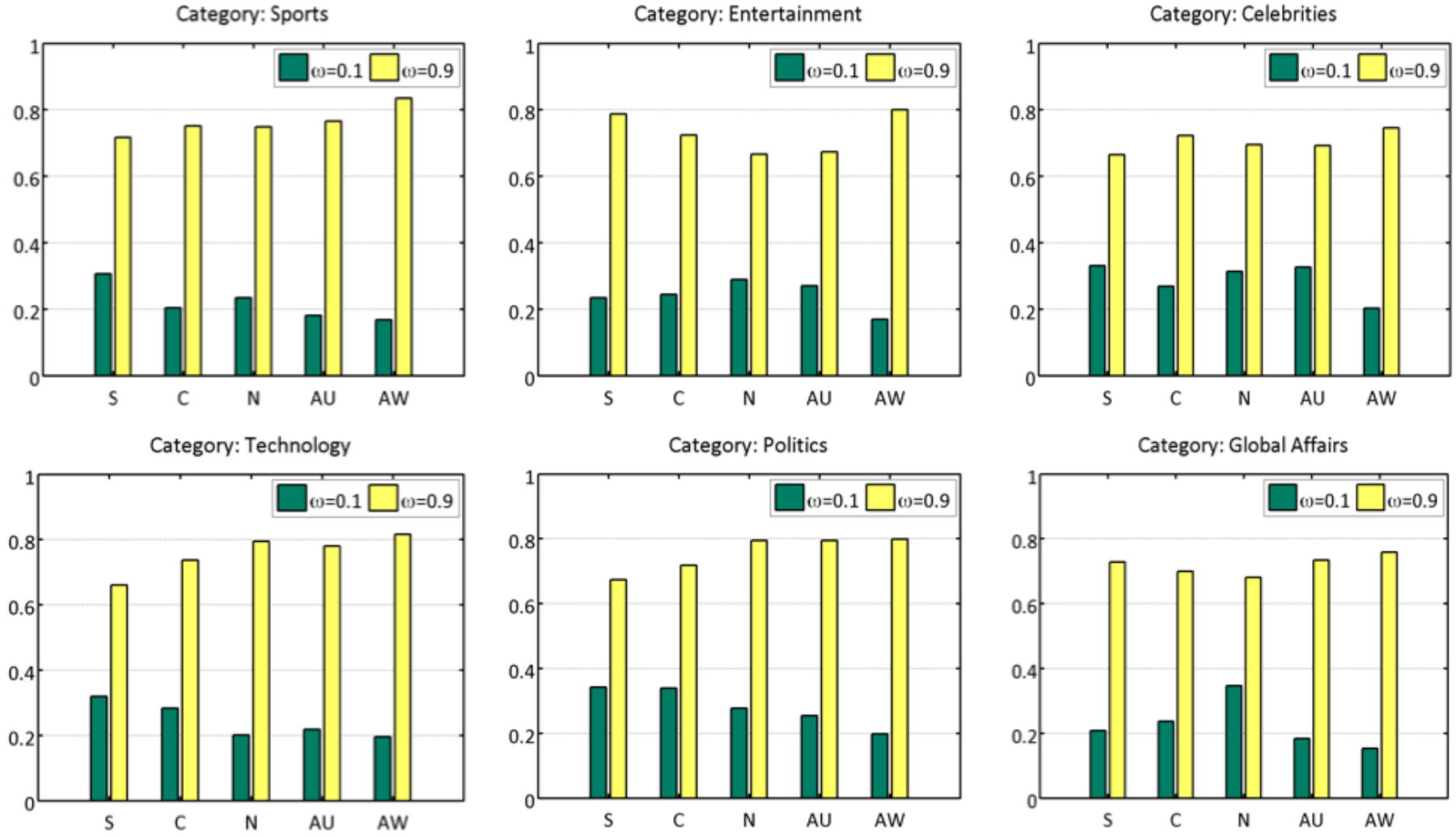






iPhone

Impact of Dimensions



S=social, C=content, N=nodal, AU=all features (unweighted), AW=all features (weighted)

How does the sampling process impact users' cognitive abilities of information consumption?

Cognitive metrics

- *Explicit Measures*. Explicit measures consisted of three 7-point Likert scale ratings made after reading each tweet set,
  - "interestingness"
  - "informativeness"
- *Implicit Measures*.
  - Subjective Duration Assessment [Czerwinski 2001] – ideally if the information presented in a tweet sample is very engaging, the participant would underestimate the time taken to go through it.
  - Recognition Memory for tweets already shown – related to encoding in the long-term memory [Sperling 1973, Smith 1979].

Part I

Please read the following sample of 10 tweets. When you are done reading, click the "Finished Reading!" button below to take a short evaluation of the tweet sample.

Topic: Oil Spill [Tweet Sample, 3 of 12]

From user, @[redacted]	Tweet: Will The Oil Spill Affect You? http://blog.expertox.com http://bit.ly/9xt5Od	Posted at: 2010-06-07 06:59:50
From user, @[redacted]:	Tweet: RT @rbndvd Blood used to be thicker than water. That was before the BP oil spill though.	Posted at: 2010-06-07 07:00:50
From user, @[redacted]	Tweet: RT @AP: AP Essay: Gulf oil spill is a reminder of why Americans have lost faith in nearly every national institution. http://bit.ly/cBcK ...	Posted at: 2010-06-07 07:01:24
From user, @[redacted]	Tweet: http://bit.ly/bpaQD2 Gulf oil spill: Containment cap working well so far, says BP	Posted at: 2010-06-06 15:07:37
From user, @[redacted]:	Tweet: RT @ScottBourne: If you find this meaningful I'd appreciate a RT - Don't Think Photography's Important? Impact of BP Oil Spill - http:// ...	Posted at: 2010-06-07 06:36:37
From user, @[redacted]:	Tweet: BP Tries To Spin Oil Spill - Watch BP's New Ad (Video) - IndyPosted http://bit.ly/c4kkYQ	Posted at: 2010-06-06 15:40:05
From user, @[redacted]	Tweet: RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a	Posted at: 2010-06-07 06:43:13
From user, @[redacted]	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://fga.me/99fc69	Posted at: 2010-06-06 18:51:51
From user, @[redacted]:	Tweet: Oil Spill: http://www.aquarianadvertising.com/info/wordpress/?p=3530	Posted at: 2010-06-07 05:53:45
From user, @[redacted]	Tweet: Oh yeah... Totally forgot about the stupid oil spill. Now I can't swim to the Bahamas lol	Posted at: 2010-06-06 20:20:56

User Study…

Part I

Please read the following sample of 10 tweets. When you are done reading, click the "Finished Reading!" button below to take a short evaluation of the tweet sample.

Topic: Oil Spill [Tweet Sample, 3 of 12]

From user, @[redacted]	Tweet: Will The Oil Spill Affect You? http://blog.expertox.com http://bit.ly/9xt5Od	Posted at: 2010-06-07 06:59:50
From user, @[redacted]:	Tweet: RT @rbndvd Blood used to be thicker than water. That was before the BP oil spill though.	Posted at: 2010-06-07 07:00:50
From user, @[redacted]	Tweet: RT @AP: AP Essay: Gulf oil spill is a reminder of why Americans have lost faith in nearly every national institution. http://bit.ly/cBcK ...	Posted at: 2010-06-07 07:01:24
From user, @[redacted]	Tweet: http://bit.ly/bpaQD2 Gulf oil spill: Containment cap working well so far, says BP	Posted at: 2010-06-06 15:07:37
From user, @[redacted]:	Tweet: RT @ScottBourne: If you find this meaningful I'd appreciate a RT - Don't Think Photography's Important? Impact of BP Oil Spill - http:// ...	Posted at: 2010-06-07 06:36:37
From user, @[redacted]:	Tweet: BP Tries To Spin Oil Spill - Watch BP's New Ad (Video) - IndyPosted http://bit.ly/c4kkYQ	Posted at: 2010-06-06 15:40:05
From user, @[redacted]	Tweet: RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a	Posted at: 2010-06-07 06:43:13
From user, @[redacted]	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://fga.me/99fc69	Posted at: 2010-06-06 18:51:51
From user, @[redacted]:	Tweet: Oil Spill: http://www.aquarianadvertising.com/info/wordpress/?p=3530	Posted at: 2010-06-07 05:53:45
From user, @[redacted]	Tweet: Oh yeah... Totally forgot about the stupid oil spill. Now I can't swim to the Bahamas lol	Posted at: 2010-06-06 20:20:56

Now please respond to the following questions below:

a. Estimate the length of time, in minutes and seconds (e.g. in the format "X min, Y sec"), you think you needed to go through the tweets.

[ ] min, [ ] sec

b. **INTERESTINGNESS:** How interesting did you find the tweets in the sample shown? In the scale below, 1 means not at all interesting, 7 means highly interesting.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

c. **DIVERSITY:** How diverse did you find the tweets in the sample shown? A diverse set of tweets would contain different sub-topics, would appear to come from different parts of the world, would contain a mix of tweets and re-tweets, etc. In the scale below, 1 means the tweets are not at all diverse, 7 means they are highly diverse.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

d. **INFORMATIVENESS:** How informative did you find the tweets in the sample shown? Note, although you'll notice that there are some repeating tweets across samples, rate the informativeness of the sample as a whole. In the scale below, 1 means the sample is not at all informative, and 7 means the sample is highly informative.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

User Study…

Part III

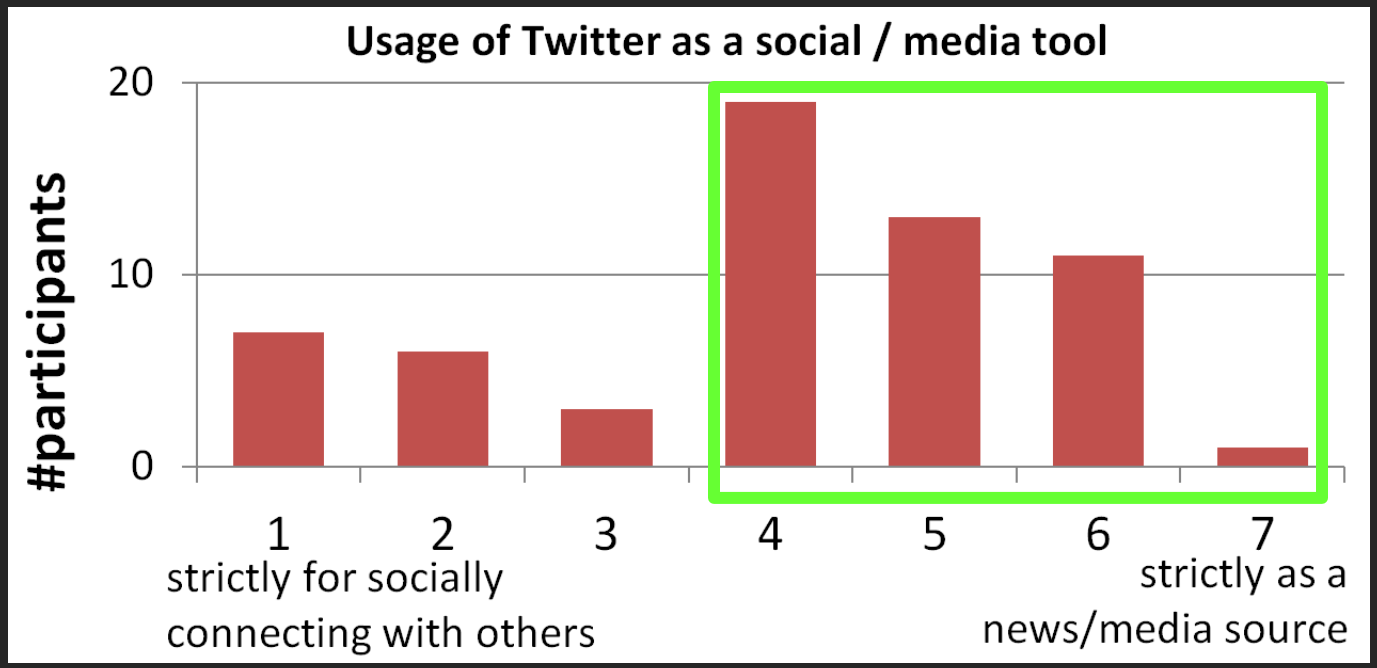
In this final part of the study you are required to go through the following 72 tweets as presented below. Some of these you would have seen before, while others you wouldn't have seen. Recognize if each of them was shown to you in any of the former pages. Each tweet has a "Yes" / "No" option: so please use your memory to recognize if you saw the tweet or not ("Yes" if you saw it, and "No" if you didn't). Good luck!

From user, @[redacted]	Tweet: RT @malloryallyce: Yo, everyone buy Dawn dish soap $1 of each bottle goes to helping the poor animals affected by the oil spill. :(	Posted at: 2010-06-06 16:59:30	Yes No
From user, @[redacted]	Tweet: RT @ElevateU: RT @PoliticalTicker: House subcommittee holds hearing on oil spill http://bit.ly/clMJ4a	Posted at: 2010-06-07 06:28:32	Yes No
From user, @[redacted]:	Tweet: I think Obama is really killing his chance of re-election with the happening and handling of the BP oil spill. Is this Obama's 9/11?	Posted at: 2010-06-07 16:18:11	Yes No
From user, @[redacted]:	Tweet: RT @nytimesscience: Pelicans, Back from Brink of Extinction, Face Threat From Oil Spill http://nyti.ms/cFGUoN	Posted at: 2010-06-07 12:55:47	Yes No
From user, @[redacted]:	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://fga.me/99fc69	Posted at: 2010-06-06 18:51:51	Yes No
From user, @[redacted]	Tweet: RT @JasonLeopold: RT @EnvironUpdates: NPR: Scientists: Dispersants Compounded Oil Spill http://bit.ly/dC0V6t Full http://n.pr/b51MvU	Posted at: 2010-06-07 02:44:50	Yes No

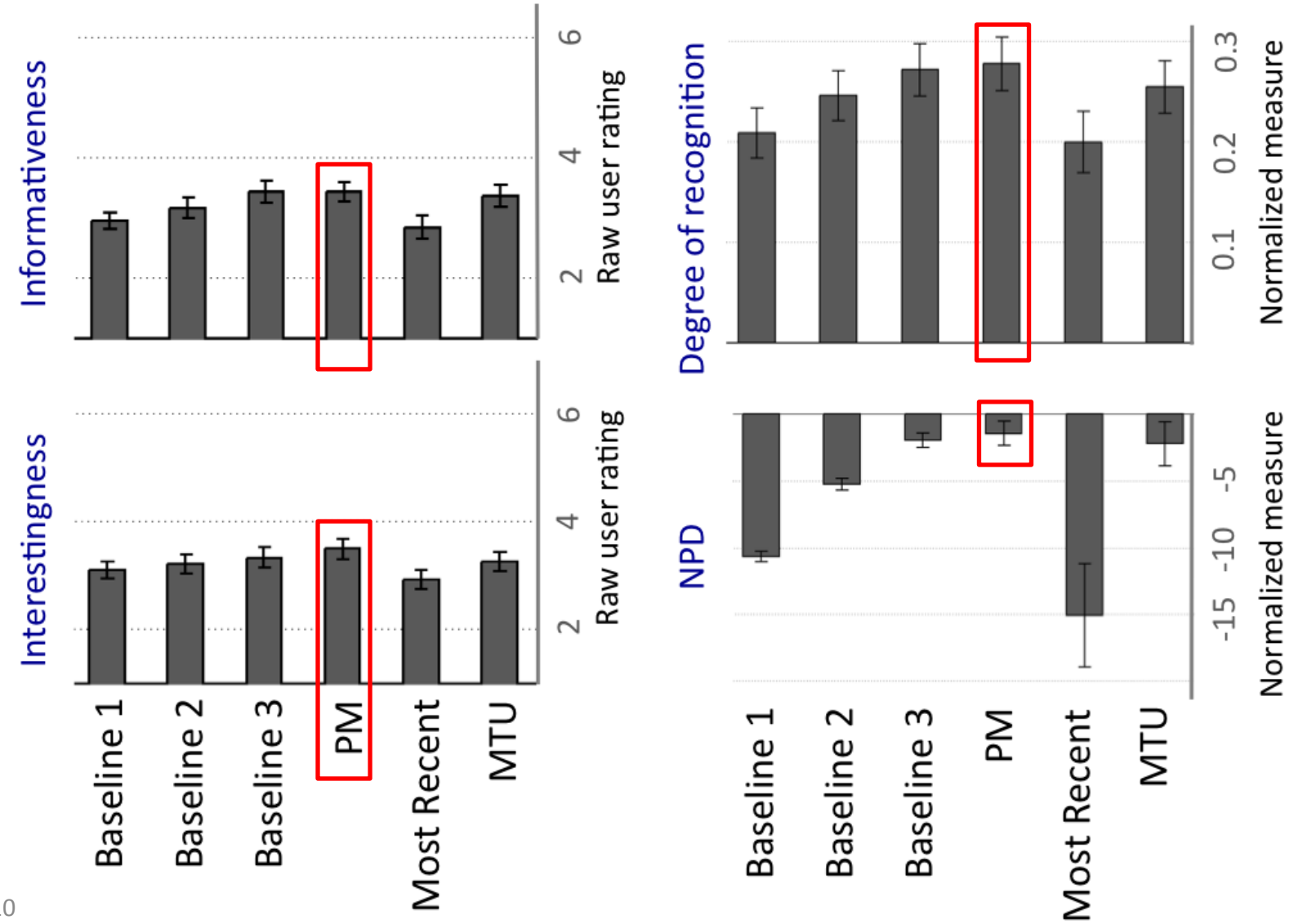
User Study…

User Study…

- 67 participants at a large organization (60% male, 40% female), median age 26 years.
- Samples on two trending topics from Twitter evaluated: “Oil spill” and “iPhone”.
- Three levels of diversity considered for the samples: 0.1, 0.6 and 0.9.



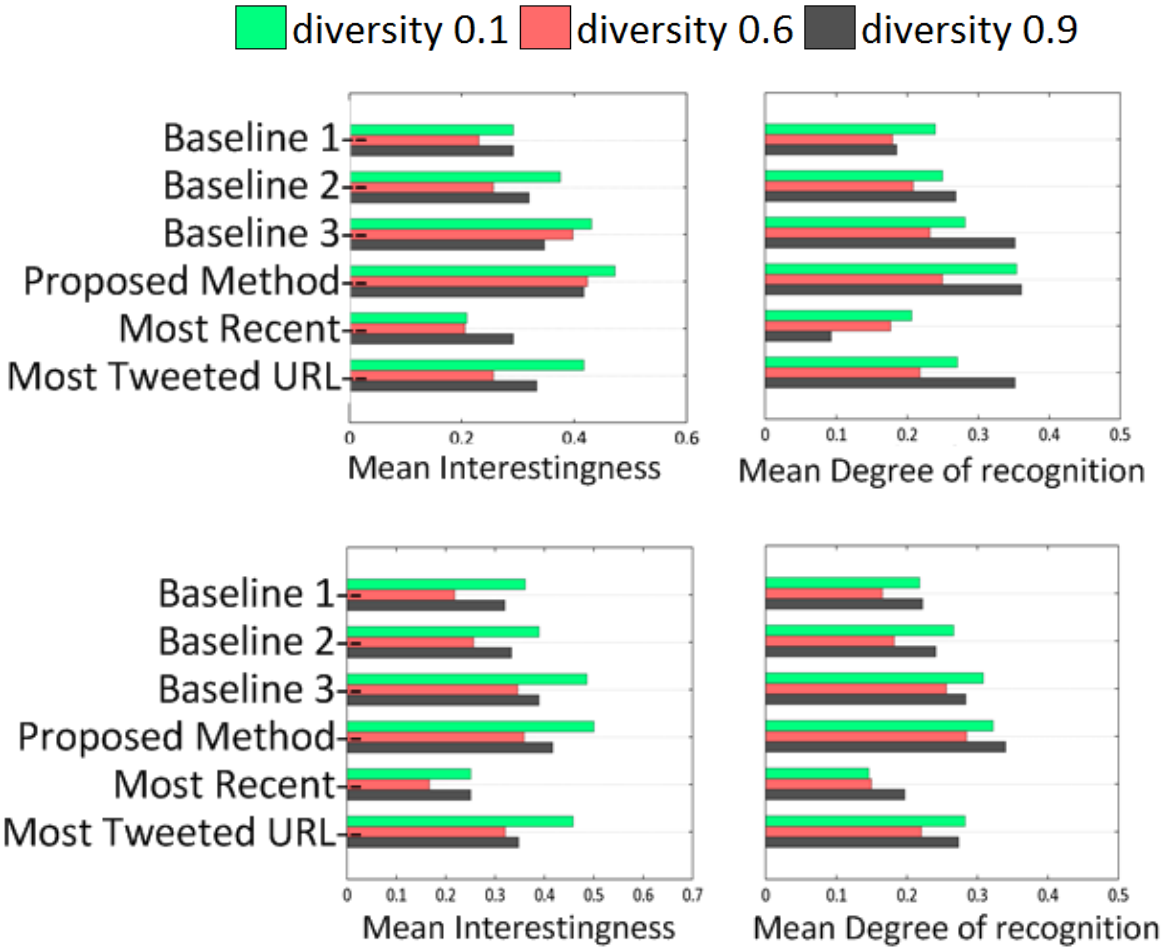
Evaluation in terms of Cognitive Metrics



What is the role of diversity in the sampling process?

Are there empirical bounds on what degrees of diversity in a sample best suit content consumption?

Diversity perception



Participant ratings on different cognitive aspects of information consumption seems to be higher for highly homogenous and highly heterogeneous information samples

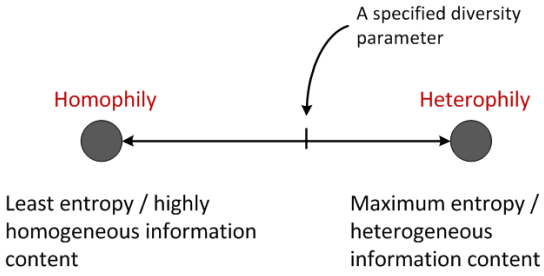
So is there more to the information space topology that can guide the sampling methodology?

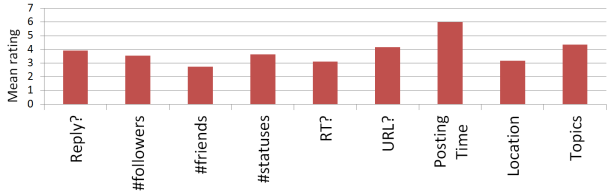
How robust are the impacts of these entropy signatures?

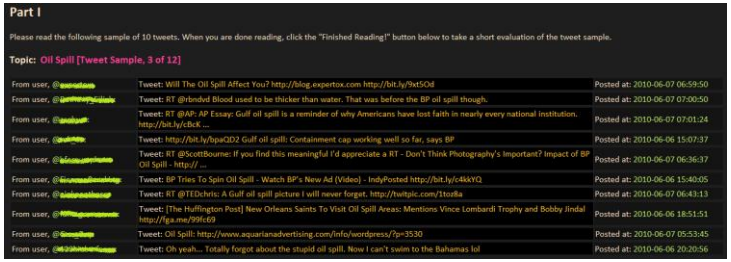
Entropy signature ⟶ Sampling?

Conclusions

- Sampling methodologies of large social information spaces that incorporate cognitive metrics of content consumption can enable the design of better content exploration interfaces.
  - Information diversity is key
  - User appear to cognitively encode information better, when presented with samples of high or low diversity
  - Our proposed sampling algorithms that incorporate cognitive metrics of content consumption perform better than straw-man versions of state-of-the-art techniques








The End
NEXT EXIT

Social networks and media are causing significant changes in our lives

Inferences about social phenomena is affected by *data quality*

Streamlining the user experience is affected by *data relevance*

And it matters completely…

Acknowledgements

- Advisor, Prof. HariSundaram, CS +AME, Arizona State University.
- Collaborator, Dr. Doree Duncan Seligmann, Avaya Labs Research.
- Collaborator, Dr. Duncan Watts, Yahoo! Research.
- Collaborator, Dr. Scott Counts, Microsoft Research.
- Collaborator, Dr. Mary Czerwinski, Microsoft Research.
- *Twitter data: "*full fire-hose" over June 2010, courtesy, Microsoft Research, Redmond.





Questions?

For details: munmun@asu.edu
Web: http://www.public.asu.edu/~mdechoud/
Twitter: @munmun10

Appendix

Prediction Tasks: Node Status/Gender

- Given feature set of structural features & mean edge weight of neighbors with attribute *i:*

$$f_i^{\tau} = \left\{ k_i^{\tau}, k_i^{(2),\tau}, k_i^{(n),\tau}, \Phi_i^{\tau}, \mathrm{X}_i^{\tau}, \eta_i^{\tau}, \omega_1 \cdot \left| N_i(a_1) \right|, \omega_2 \cdot \left| N_i(a_2) \right|, \ldots, \omega_q \cdot \left| N_i(a_q) \right| \right\}$$

  where ω_j gives the mean edge weight of u_i *with* respect to the neighbors having attribute value j $(1 \le j \le q)$ and $N_i(a_j)$ *is the subset of i's neighbors whose attribute* value is j
  - Also consider an unweighted version with all ω_j=1
- Split into training (90%) and test (10%) sets
- Use SVM (Support vector machine based attribute prediction) with Gaussian RBF kernel, learn parameters & kernel width with *k*-fold cross-validation (*k*=10 in this work)

Prediction Tasks: Future Communication

- To predict activity of a user u_i at time t_{m+1}, we use a similar feature-based representation of u_i in the network $G(\tau)$, i.e.
  - the structural features
  - the mean weighted activities of her neighbors from time t_0 to t_m
  - we augment the feature space by using u_i's communication from t_0 to t_m
- We fit a linear model of communication activity as a function of the node level features $\mathbf{F}^{\tau}_{0:m}$:

  $A_m = \beta^{\tau}_{0:m} \cdot \mathbf{F}^{\tau}_{0:m} + \varepsilon^{\tau}_{0:m}$, where $\varepsilon^{\tau}_{0:m}$ is additive noise.

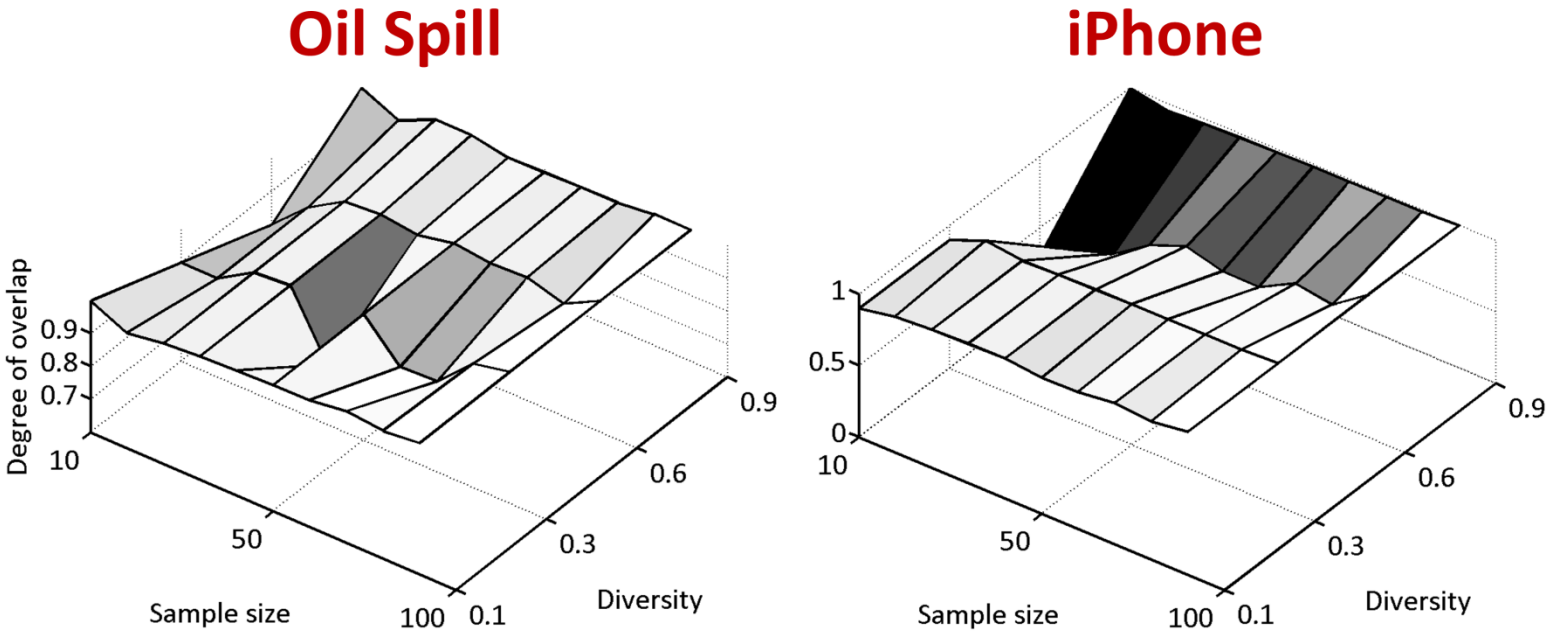
- The best-fit coefficients $\beta^{\tau}_{0:m}$ are used along with the feature vector at t_{m+1}, to predict future node activity given as $A'_{m+1} \in \mathbf{R}^{1\times|V|}$
- **University dataset:**
  - we divide the data over the span of two years into the six different semesters and regress over the first five semesters to predict the activity at the sixth semester
- **Enron dataset:**
  - we divide the span of activity over four years (1998-2002) into time intervals of t_i = 3 months each

Prediction Tasks: Community Detection

- Fit a stochastic block model to $G(\tau)$ using variationalBayes inference [Hofman et al. 2008]
- ***Method:***
  - Assume each node u_i belongs to one of the Z latent groups/"blocks" (or school assignments) , given as $z_i$with probability π_μ, μ=1,2,…,Z
  - If the nodes $u_i$and u_j are in the same group (z_i=z_j), an edge exists between them with probability ϑ_+; if they are in different groups ($z_i \neq z_j$), an edge exists between them with with probability ϑ_-
  - Given only the observed edges $e_{ij} \in E_s$ in the graph $G(\tau)$, distributions over the group assignments $p(z_i)$ are inferred via variational Bayesian inference
- Compare soft assignments to actual school affiliation using normalized mutual information
- In our experiments, Z=5 for the University dataset

Robustness of sampling method

- Robustness of proposed sampling method across multiple iterations.
  - We show the degree of overlap of tweets corresponding to samples that are generated across iterations. The overlap values are shown for various sample sizes as well as three diversity parameter levels.



Scalability

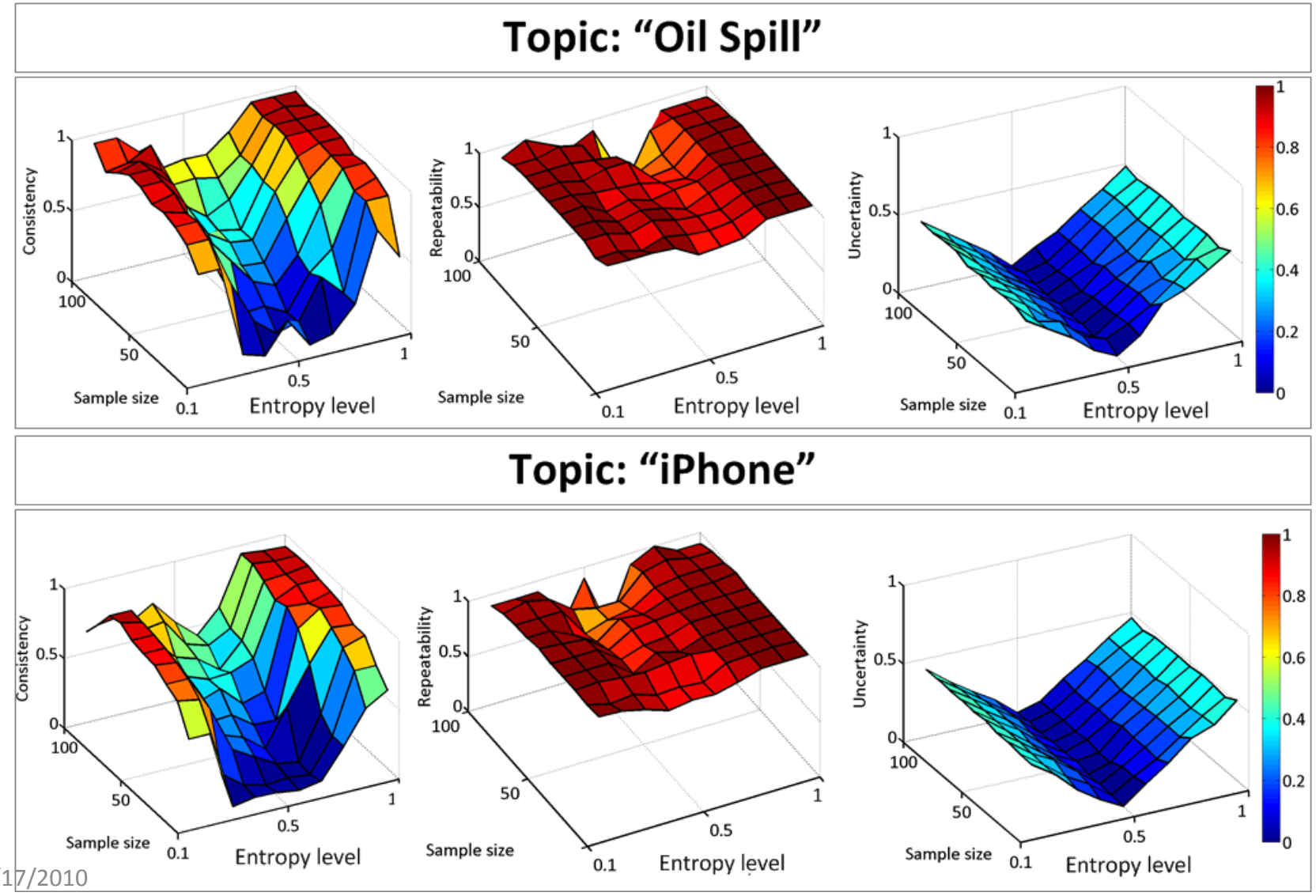
List of 30 trending topics from Twitter that were used for studying the robustness of our proposed sampling method. Broad thematic categories (hand-labeled) are indicated to indicate a wide span of topics.

Type	Trending Topics
Sports	NBA, Vuvuzela, #worldcup, Lakers, Suns
Entertainment	Star Trek, Harry Potter, New Moon, Twilight, American Idol, Inception
Celebrities	Lady Gaga, Michael Jackson, Justin Bieber, Lindsay Lohan
Technology	Tweetdeck, iPad, Snow Leopard, iPhone, Apple, At&t, Google wave, Motorola
Politics	BarackObama, McCain, Afghanistan
Global Affairs	H1N1, Haiti, Oil Spill

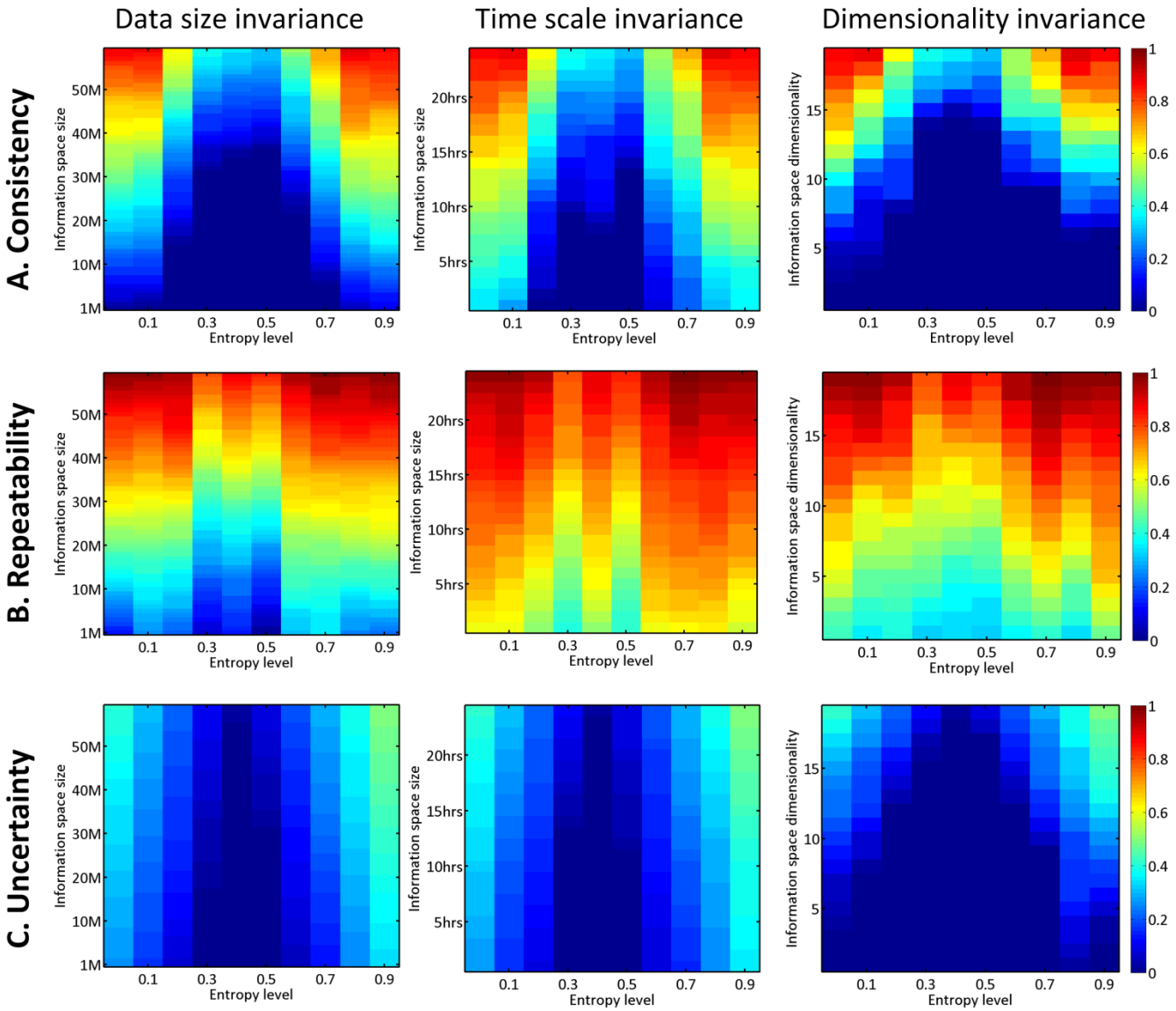
Statistical significance of performance of our proposed method across all the 30 trending topics. Performance is evaluated in terms of the L_1-norm distance between the entropy of the samples generated, and the desired diversity parameter values: 0.1 through 0.9, in increments of 0.1. High *p*-value indicates that the dierences across topics are not significant, i.e., our method is consistent across topics.

SS	df	MS	*F*-statistic	*p*-value
0.18254	29	0.00629	0.85	0.6864

Characteristics of entropy signatures in sampling process



Invariance property of entropy signatures



Correlation of signature characteristics with sampling interval